Summer 2002

EDUCATION next

A JOURNAL OF OPINION AND RESEARCH

Up Against the Wall

Will the Supreme Court end the voucher debate?

JOSEPH VITERITTI • PAUL E. PETERSON • WILLIAM HOWELL

IN THIS ISSUE:

Sizing Up Value-Added Analysis, New American Schools, The Power of Peers, The Bush-Kennedy Courtship, Education at Ground Zero, Reading Richard Rothstein

CONTRIBUTORS: DALE BALLOU, JOEL BEST, SIOBHAN GORMAN, JAY GREENE, CAROLINE HOXBY, MICHAEL KIRST, DONALD McADAMS, JEFFREY MIREL, DAVID MURRAY, ANITA SUMMERS

www.educationnext.org / $7.00

HOOVER DIGEST

Informative. Provocative. Insightful.

WORLD WAR
HOOVER DIGEST
Michael Barone
Politics after September 11
AFTERMATH
Daniel Brumberg
Stalemate in Tehran

Turn to the quarterly *Hoover Digest* for lively and compelling writing on politics, economics, history, and culture from the Hoover Institution, one of the nation's preeminent think tanks.

A Free Issue with No Obligation! Call now and receive a free copy of the latest issue. If you like what you see, you can subscribe at the special introductory rate of $20.

TO RECEIVE YOUR FREE ISSUE, CALL 800-935-2882.

Hoover Institution, Stanford University, Stanford, California 94305-6010 fax: 650-723-8626

www.hooverdigest.org

VOLUME 2, NUMBER 2

EDUCATION next

A JOURNAL OF OPINION AND RESEARCH

9

24

36

forum

9 Accountability Gains

Are we measuring achievement gains accurately enough?

Sizing Up Test Scores by DALE BALLOU *Expert Measures* by ANITA A. SUMMERS

The Business Model by JAY P. GREENE *Enemy of the Good* by DONALD R. McADAMS

features

24 Vouchers on Trial

Will the Supreme Court's decision in Zelman *end the debate?* by JOSEPH P. VITERITTI

Inside the Courtroom by PAUL E. PETERSON

36 Bipartisan Schoolmates

President Bush forges a consensus on federal education policy by SIOBHAN GORMAN

44 Swing State

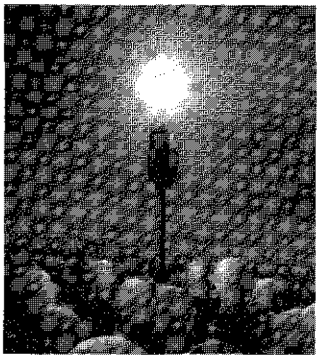
The downs and ups of accountability in California by MICHAEL W. KIRST

50 Monster Hype

School violence, the media's phantom epidemic by JOEL BEST

research

56 The Power of Peers
How does the makeup of a classroom influence achievement?
by CAROLINE M. HOXBY



56

64 Unrequited Promise
Tracing the journey of New American Schools
by JEFFREY MIREL

check the facts

73 Waiting for Utopia
Richard Rothstein, the New York Times *education columnist*
CHECKED by DAVID W. MURRAY

New York Times

73

from the editors

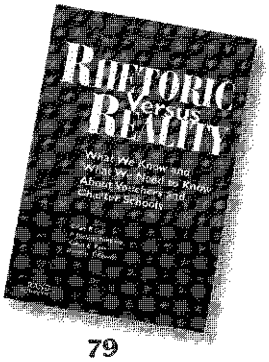
5 Two Steps Forward?

correspondence

6 *NCATE responds*

book review

79 Data Vacuum
Only larger voucher experiments will yield answers
REVIEW ESSAY by WILLIAM G. HOWELL


RHETORIC Versus REALITY
What We Know and What We Need to Know About Vouchers and Charter Schools

79

84 Revolution at the Margins
The Impact of Competition on Urban School Systems
by Frederick M. Hess
REVIEWED by EDWARD B. FISKE & JOHN GARDNER

education matters to me

88 In the Shadow of Terror
by LAURA KRUG

88

COVER ILLUSTRATION *by Michael Gibbs*

VOLUME 2, NUMBER 2
EDUCATION next
A JOURNAL OF OPINION AND RESEARCH

Editor-in-Chief	Paul E. Peterson
Senior Editor	Chester E. Finn Jr.
Executive Editors	Frederick M. Hess Marci Kanstoroom
Research Editor	Martin West
Articles Editor	Tyce Palmaffy
Managing Editor	Carol Peterson
Design Director	John Sizing
Consulting Editors	Patricia Baker, Jay P. Greene
Editorial Staff	Samuel Abrams, Tammy Frisby, Susan Pasternack, Deborah Styles
Web Editor	Bridgitt Ahern
Circulation	Jennifer Presley
Editorial Board	John E. Chubb, Williamson Evers, Eric A. Hanushek, Paul Hill, E. D. Hirsch Jr., Caroline M. Hoxby, Terry M. Moe, Diane Ravitch, Herbert Walberg

Publisher	Hoover Institution, Leland Stanford Junior University
Additional Sponsoring Institutions	Program on Education Policy and Governance, Harvard University Thomas B. Fordham Foundation Manhattan Institute for Policy Research

Subscriptions One-year subscription (four issues), $20 in the United States; $30 in Canada; $40 elsewhere. To subscribe or to submit change of address information, call us at (650) 723-3373; fax us at (650) 723-8626; visit us at www.educationnext.org; or email us at hooverpress@hoover.stanford.edu. You may also write to *Education Next* Subscriptions, Hoover Institution, Stanford University, Stanford, CA 94305-6010. Cover price is $7/issue. ISSN 1532-5148.

Advertising For information on advertising, contact Thomas Polseno, Program on Education Policy and Governance, Kennedy School of Government, Harvard University, Cambridge, MA 02138, (617) 495-6954.

Letters We welcome letters about our contents. Longer versions of some letters may be found at www.educationnext.org. Write to Letters to the Editor, *Education Next*, 226 Littauer North Yard, 1875 Cambridge Street, Cambridge, MA 02138; fax to (617) 496-1507; or email Editor_EM@latte.harvard.edu.

Editorial Queries We consider unsolicited contributions. Send manuscript submissions as email attachments to Editor_EM@latte.harvard.edu. Pieces sent by regular mail go to Carol Peterson, *Education Next*, 226 Littauer North Yard, 1875 Cambridge Street, Cambridge, MA 02138.

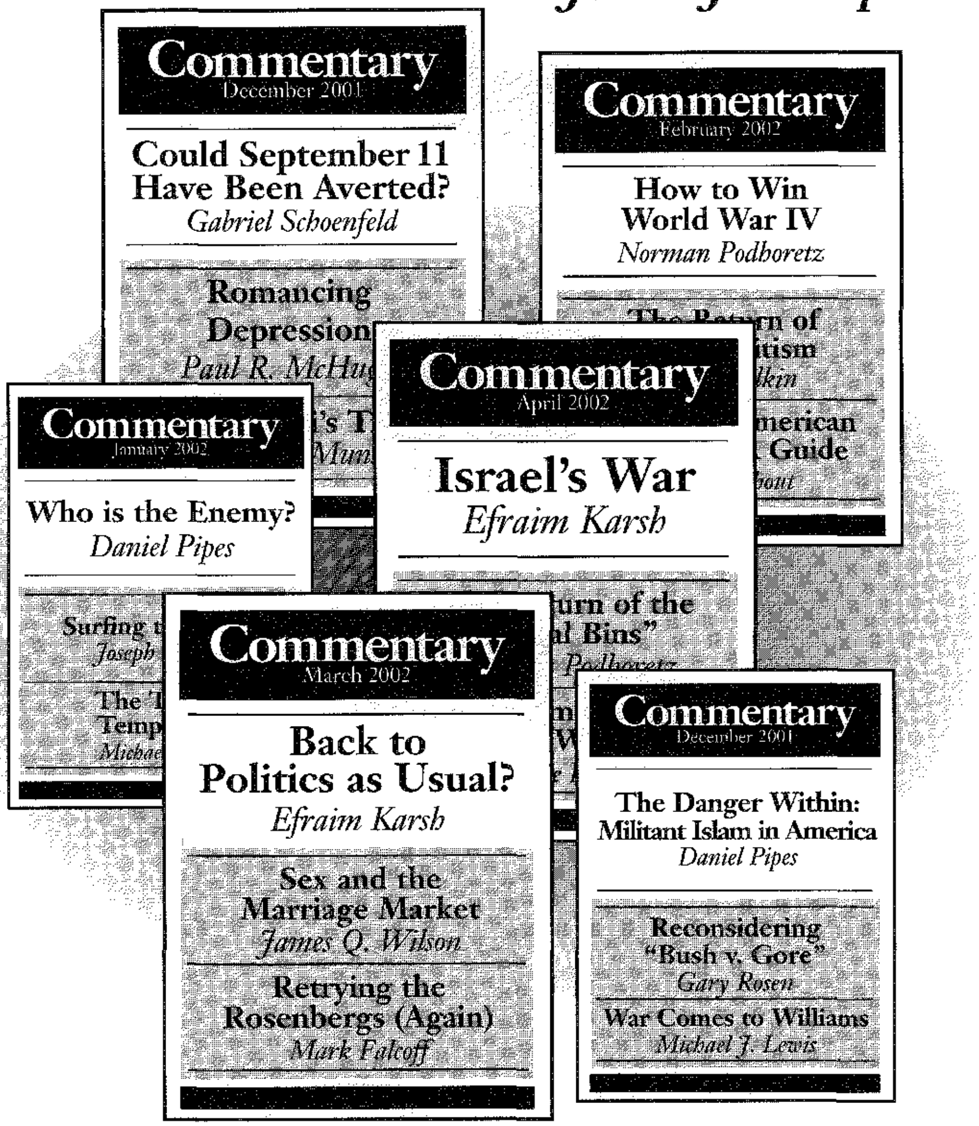
Opinions expressed in *Education Next* are those of the authors and do not necessarily reflect those of the Editorial Board of *Education Next*, the Board of Trustees of Leland Stanford Junior University, or the sponsors of *Education Next*.

Visit us at http://www.educationnext.org, where unabridged essays can be found.

© by the Board of Trustees of Leland Stanford Junior University, 2002

"Very, very important to the life of the United States, to the West, and, I am convinced, to freedom."

—*Jeane J. Kirkpatrick*



Commentary
December 2001
Could September 11 Have Been Averted?
Gabriel Schoenfeld
Romancing Depression
Paul R. McHu

Commentary
February 2002
How to Win World War IV
Norman Podhoretz

Commentary
January 2002
Who is the Enemy?
Daniel Pipes

Commentary
April 2002
Israel's War
Efraim Karsh

Commentary
March 2002
Back to Politics as Usual?
Efraim Karsh
Sex and the Marriage Market
James Q. Wilson
Retrying the Rosenbergs (Again)
Mark Falcoff

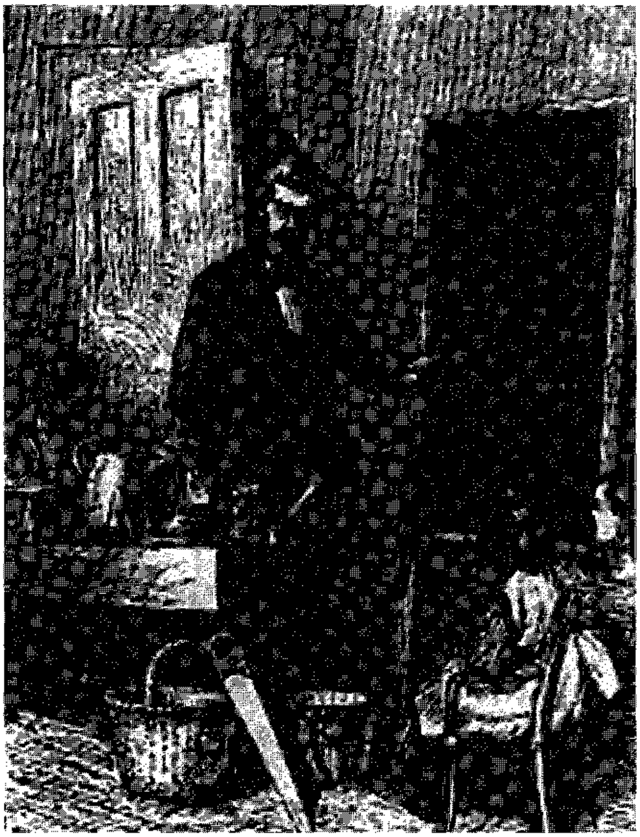
Commentary
December 2001
The Danger Within: Militant Islam in America
Daniel Pipes
Reconsidering "Bush v. Gore"
Gary Rosen
War Comes to Williams
Michael J. Lewis

Commentary

Sound thinking for the 21st century.

$45.00 for 11 issues • Subscribe today! • $79.00 for 22 issues

Call 1 800 829-6270



penn

1-800-326-9180

state

www.psupress.org

press

JGE: THE JOURNAL OF GENERAL EDUCATION

Marilyn Amey, Editor
Elizabeth Jones, Sr. Associate Editor
David Guthrie and Elizabeth Larsen, Associate Editors

"As a member of the university council working on a revision of undergraduate general education requirements, I found the articles in JGE of particular relevance. The timely studies encouraged by the editorial board promise to create a forum for lively exchange on educational issues."

—Julia Herschensohn, University of Washington

For faculty, administrators, and policymakers, JGE is the professional forum for discussing issues in general education today. JGE addresses the general education concerns of community colleges, four-year colleges, universities, and state systems. Along with perceptive essays on the role of general education today, JGE features articles on:

- Innovative methods in teaching and assessment
- Profiles of exemplary general education programs
- Case studies of successful curriculum development efforts
- Reviews of books and monographs related to general education

Annual Personal Subscriptions: $27.00; Outside USA $35.00
Annual Institutional Subscriptions: $47.00; Outside USA $54.50
Student Subscriptions: $20.00; Outside USA $28.00
Air Mail: Add $36.00 per volume
Terms: Prepayment is required (Mastercard, Visa, check, or money order)
Journals Subscription Office: 814-863-5992

Project MUSE® **http://muse.jhu.edu/journals/jge/**

Two Steps Forward?

2002, a year full of promise in education reform

Although September 11 briefly arrested the nation's work on domestic issues, 2002 is still shaping up as a significant year for education reformers.

When President Bush affixed his signature to the No Child Left Behind Act on January 8, 2002, he arguably brought to life the most important piece of federal education legislation since 1965. For the first time, Washington would now require all states, in exchange for federal dollars, to demonstrate measurable progress in student learning. As Siobhan Gorman shows, in a domestic version of "Nixon goes to Beijing," it was a Republican administration that decisively expanded the federal role. Congress and the president have placed a new and only partially funded mandate on state and local governments—the kind of measure that congressional Republicans, just a few years ago, were swearing never to do again. Resistance to this new mandate came mainly from those on the Democratic side of the aisle. These one-time proponents of a strong national role in education managed to water down the law's accountability provisions to the point where it remains unclear whether the legislation's promise will be realized. Michael Kirst's essay on the ups and downs of California's accountability efforts should give pause to any who take for granted that the new federal law will be implemented.

In a domestic version of "Nixon goes to Beijing," it was a Republican administration that decisively expanded the federal role.

Many of those implementation issues involve the arcane matter of measuring progress. Ordinarily, yardstick questions are left for discussion in the halls of statistics departments, but the topic has become so pressing that we have devoted this issue's **Forum** to "value-added" measurement. The theory seems straightforward: determine how much a student learned in a given year by subtracting from his or her most recent test scores the results of the previous year's tests. Would that it were so simple. Measuring a student's performance twice is an invitation to make two mistakes. Classroom noise, illness, confusing instructions by an inexperienced teacher, too much help from the teacher-coach—these and many other elements can confound the measurement. All the forum participants acknowledge the problem, though they disagree about its implications for testing policy.

The year is shaping up to be momentous for the school choice movement as well, though again the long-term implications remain misty. On February 20, the U.S. Supreme Court heard oral argument regarding the constitutionality of the Cleveland voucher program. That the justices allowed 80 minutes for argument, instead of the usual 60, was just one measure of the case's significance. It was the questioning of Justice Sandra Day O'Connor, widely regarded as the swing vote in this case, that received the closest attention. To some veteran observers, the case was decided when she commented: "If anything, [the law is] skewed against the religious schools." Still, no one will know for sure until the court hands down its decision, probably in May or June of this year. Joseph Viteritti's **Feature** tells us that even a decision allowing the Cleveland program to proceed will be followed by much litigation over the many state-created "walls of separation." Meanwhile, as William Howell explains in his review of recent overviews of school voucher research, we can look forward to a continuing flow of "objective," "reasonable," and "comprehensive" summaries of recent research, each written authoritatively from a particular ideological perspective.

The issue contains still more. Joel Best shows how statistics can be used to distort the public's perception of the amount of violence and bullying in our schools. Caroline Hoxby's **Research** is a creative test of the widely accepted proposition that students' peers affect how much they learn. Predictably, they do—but not always in the ways you might have expected. Jeffrey Mirel tells the fascinating story of how New American Schools evolved from an ambitious plan to foster education change into a Beltway enterprise dependent largely on federal grants. In **Check the Facts,** David Murray explains how One Big Idea shapes the commentary of Richard Rothstein, whose writings appear on the news pages of America's most influential paper.

-THE EDITORS

MISSION STATEMENT In the stormy seas of school reform, this journal will steer a steady course, presenting the facts as best they can be determined, giving voice (without fear or favor) to worthy research, sound ideas, and responsible arguments. Bold change is needed in American K-12 education, but *Education Next* partakes of no program, campaign, or ideology. It goes where the evidence points.

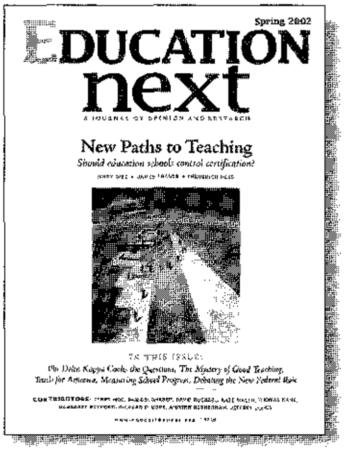
Quick fix

Margaret Raymond and Stephen Fletcher's findings ("Teach for America," *Research*, Spring 2002) from their initial evaluation of Teach for America (TFA) are not too surprising, given the makeup of TFA recruits and the teachers with whom they are being compared.

They find that TFA recruits in Houston are "at least as effective as other teachers in the district." However, the researchers readily admit to the limits of their research, in that many of the other teachers in the Houston district have gone through quick-fix alternative certification programs.

TFA recruits "are a select group of college graduates, culled from the finest universities and often performing near the top of their class," write Raymond and Fletcher. Individuals with these qualifications are not the norm in alternative-certification recruitment. In Prince George's County, Maryland, hairdressers and cabdrivers answered ads for teachers in 2000 and 2001. Usually, alternative routes attract those who have not thought about their career choices and are looking for something to do or are unemployed. In fact, during the recession after September 11, 2001, applications for teaching shot up in various locations across the country. Most alternative-certification applicants usually end up in the classroom because something better has not turned up. This is a far cry from the TFA recruits, with their Ivy League degrees and 3.4 GPAs.

During the past decade, TFA has placed 7,000 recruits in classrooms—but only 2,000 to 3,000 of those are still in the classroom. Retention is lower than among regularly trained teachers, creating constant turnover in those schools that choose TFA recruits. This may give TFA graduates a realistic picture of the problems in education (indeed, they're contributing to some of them, like teacher turnover in urban schools), and they may advocate for solutions as they move on to their "real" careers.



Spring 2002

EDUCATION next

New Paths to Teaching

Should education schools control certification?

However, TFA is not an answer to the teacher shortage in America. America needs almost 200,000 new teachers each year, or about 2 million in the next decade. The select group of top college graduates that includes TFA members will not stay for long in jobs that are at the bottom of the pay and perk scale—jobs that in many urban and suburban locations make them eligible for low-income housing. The vast majority did not attend Ivy League schools to earn less in a year than it cost for nine months of their undergraduate education.

If the salaries and working conditions of teachers were raised to a level commensurate with those of other college graduates, the profession would begin to attract more of the best and brightest into teacher preparation programs.

Jane Leibbrand
National Council for Accreditation
of Teacher Education
Washington, D.C.

No alternative

David Ruenzel's article ("Tortuous Routes," *Feature*, Spring 2002) accurately describes a major deficiency of most alternative routes to teacher certification: Because they reflect traditional assumptions about teaching, they are no more effective than the unworkable systems they are intended to help fix.

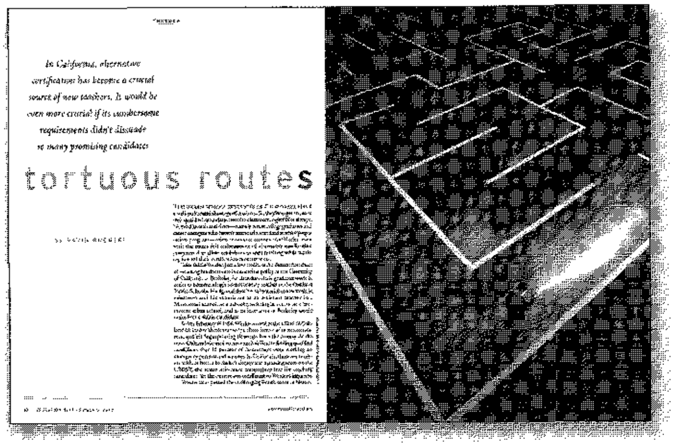
Those systems assume that teacher quality is best attained when the state heavily regulates employment, requiring teachers to take numerous education courses before they can be considered for a teaching job. The problem is that because teaching is not a science, the technical knowledge base is thin. Therefore, many education courses lack academic substance and none is highly predictive of success. Yet they dilute prospective teachers' undergraduate education by displacing academic courses. Since many top college students thus avoid education degrees, teacher-preparation programs often enroll weaker students to meet school demand. Still, education graduates do not provide a reliable supply of new teachers. Forced to commit in their teens in order to take the required courses, some change their minds by graduation and do not enter teaching, while many leave their jobs within a few years.

Because the traditional approach is unworkable, states have always maintained other pathways. However, these long-standing alternatives merely eschew the enforcement of certification requirements rather than questioning their necessity or effectiveness. If districts claim to be unable to find certified teachers, the state declares an "emergency" and authorizes the employment of uncertified people, who then begin after the fact to meet requirements that are supposed to be legal prerequisites. On the one hand, emergency policies have always barred districts from hiring talented people with degrees in academic subjects when a mediocre yet certified person was available. On the other hand, districts have used shortages to rationalize the employment of people who have not studied and do not know the subjects they will teach.

In 1983, New Jersey challenged the traditional approach at its roots. The state asserted that most undergraduate education courses are not useful and trimmed all but three from its traditional preparation programs, thus strengthening teaching degrees and making them more attractive to top college students. The state then gave *all* candidates with degrees in academic subjects the option of taking the three courses during their initial year of employment.

New Jersey's policy reflects a belief that it makes most sense to rely on the circumstantial judgments of principals, who happen to be licensed by the state to assess teaching capability. Teacher quality is achieved by expanding job opportunities, intensifying competition, and challenging school leaders to embrace responsibility for teacher selection and induction.

During the past 16 years, more than 10,000 new teachers have completed New Jersey's alternative route. Most of the state's 600 school districts—large and small, urban and suburban, wealthy and poor—have hired them. In any given year, up to 40 percent of all new teachers are alternative routers. A five-year pilot study showed that the initial 2,000 had higher test scores than education graduates, were more likely to hold advanced degrees, and more often had previous experience working with children. They were also three times more likely to be minorities and four times less likely to leave their teaching jobs during the first year.



In California, alternative certification has become a crucial source of new teachers. It would be even more crucial if its cumbersome requirements didn't dissuade so many promising candidates

tortuous routes

Most alternative routes to certification are hamstrung by the fact that they reflect traditional assumptions about teaching.

Although other states responded to such evidence by adopting alternate routes, many deserve the criticisms that Ruenzel expresses. The problem is that most of these alternatives were merely tacked onto traditional certification policies that remain unchallenged.

Some alternate routes inexplicably waive the same education courses that the state continues to require in traditional preparation programs. This causes school administrators to shy away from nontraditional candidates because it portrays them as less well "trained." Other states require alternate candidates to complete the same courses that are required in undergraduate preparation programs. This produces an alternate route that is unnecessarily cumbersome, a discouragement to candidates and employers alike. Still other states model their alternate routes after the old emergency credential policy, restricting their use to instances of teacher shortage.

The failure to challenge the traditional approach is sometimes the result of legitimate philosophical differences. More often, though, it results from political lobbying by interest groups that want to use government regulation to advance their professional image and that otherwise benefit from a tight and monopolistically controlled job market.

Leo Klagholz
Former New Jersey Commissioner of Education
Pomona, New Jersey

Asking the wrong questions

Florida's A+ program is one of many education programs being closely watched in education and policy circles. Because it includes vouchers and test-based accountability, the program has received national attention as a potential way to improve public education, and especially "failing schools." The prominence of the program has increased the stakes in research—and the need for careful analysis.

Jay Greene's analysis of the A+ program ("The Looming Shadow," *Research*, Winter 2001) provides results that are incomplete at best and probably quite misleading.

First, Greene assumes that the effects of the A+ program on school performance are due entirely to the pressure of facing vouchers. Our work, however, indicates that much of the effect came simply from the public embarrassment of a school's receiving a low grade.

Second, Greene reports only raw gains in test scores, rather than effect sizes, which makes it impossible to determine whether the effects are large or small. This is surprising because estimates of the effect sizes are easy to obtain, simply by dividing the raw score gains by the student-level standard deviation in score levels. According to the published work of Gregory Camilli and Katrina Bulkley of Rutgers University, Greene's raw score gain in reading translates into an effect size of just 0.04—equivalent to increasing a student's test score from the 50th to the 51st percentile. This effect is very small even by the standard Greene proposes, and grows smaller still if adjusted for the separate effect of school ratings.

Third, Greene performs a kind of cost-effectiveness analysis: Do vouchers produce gains at lower cost than other education reforms? Economists are trained to answer such questions by focusing on the economic costs. Greene's analysis has some technical problems, but the greater flaw is that he has asked the wrong question altogether. Those who follow the debate on vouchers closely know that the critique of vouchers has little to do with the effects of vouchers on test scores and monetary costs. It involves the possibility that vouchers will create a hierarchy of schools based on income and race, that increased choices for some may mean decreased freedoms for others, and that the roles of parents, teachers, and citizens in the education process will be changed for the worse.

Weighing these issues is the responsibility of policymakers, not economists and education researchers. However, researchers are responsible for helping to frame the questions properly. In this case, the key questions appear to be: How large do the test-score gains have to be to justify serious consideration in the debate about vouchers, equity, and social values? Is a 0.04 standard deviation gain large enough to be considered important? Also, what do these small gains suggest about the performance of the current education system? These are the questions that researchers should be putting to policymakers.

MARTIN CARNOY
Stanford University

DOUG HARRIS
Economic Policy Institute

"How large do test-score gains have to be to justify serious consideration in the voucher debate?"

Jay P. Greene responds: The degree to which Martin Carnoy and Doug Harris agree with me, at least implicitly, is remarkable. First, Carnoy and Harris do not take issue with the fact that test scores in Florida increased at schools that faced the prospect of vouchers under the A+ program. Second, they do not dispute my analysis comparing the results of the high-stakes state test and the low-stakes SAT 9, which found that the exceptional test-score gains experienced by schools facing the prospect of vouchers represented real increases in student achievement, not a manipulation of the state testing system or a result of teaching to the state test.

The only issues in dispute here are the extent to which vouchers as a sanction for chronic failure are uniquely capable of producing gains and whether those gains are large or not. Carnoy and Harris suggest that perhaps the threat of reconstitution or the stigma of being labeled a failure can motivate schools to improve. It is certainly plausible that various sanctions could be effective. My position is simply that the evidence from Florida shows that improvements occurred when vouchers were the sanction. We do not have evidence of similar quality from other states demonstrating that reconstitution or embarrassment is equally effective. Given the available evidence, we should operate under the working policy conclusion that vouchers are an effective sanction and that the effectiveness of other sanctions is unknown.

Our disagreement over the magnitude of the improvement experienced by schools facing the prospect of vouchers is more substantial, but even this disagreement is only a difference of degree, not kind. Carnoy and Harris rely upon an analysis by Camilli and Bulkley to claim that the gain experienced by schools facing the threat of vouchers was a modest one. Camilli and Bulkley compute the gains as units of standard deviations for individual students, even though we only know the variation in test scores on the school level and do not know the variation at the individual student level. They get around this by assuming a certain relationship between school-level variation and student-level variation in test score results. In other words, they are guessing, even if it is informed guessing. If their guess is off, gains as measured in terms of standard deviations are larger than they claim.

Rather than guess about what is not known, I prefer to report the gains in terms of a unit that is known and that everyone can understand—dollars. I calculate that to achieve the same test-score gain in math achieved by the prospect of vouchers, Florida would need to increase per-pupil spending by $3,484. Even the smaller gain produced in reading scores would require an increase of $888 per pupil to accomplish the same result. By any normal accounting, these gains are not small.

Even if Carnoy and Harris were correct in claiming that the gain produced in reading scores only yields an average improvement of 1 percentile point, such a gain is not trivial over a one-year period. And the larger gains achieved in math (even assuming their standard deviation estimates are correct) would be more substantial.

Given the difficulty policymakers have historically experienced in trying to devise any programs that produce large-scale, demonstrable gains in student achievement, the improvements realized by the A+ choice and accountability program in Florida suggest a very promising avenue for reform efforts.

forum

Accountability Gains

Will value-added analysis make performance-based evaluations a viable tool for school improvement?

THE PROBLEM

Family background decisively shapes student achievement. As a result, simplistic accountability systems that dole out rewards or sanctions based on test scores run the risk of punishing schools and teachers for problems beyond their control.

THE SOLUTION, PERHAPS

Attempt to measure the achievement gains that a school or teacher elicits by subtracting their latest test scores from the previous year's. These gains are less susceptible to the home influences than the simple level of achievement at one point in time. Thus, by measuring gains, we can pinpoint the "value" that a school has "added" to its students' educational experience.

THE DEBATE

Value-added analysis, while a promising innovation, suffers from various statistical shortcomings. Errors in measurement could lead to schools' and teachers' being rewarded or sanctioned wrongfully. The question is whether the benefits of value-added analysis outweigh the flaws.

10
Dale Ballou *details the pitfalls*

16
Anita A. Summers *trusts the testing experts*

20
Jay P. Greene *looks to the practices of private firms*

23
Donald R. McAdams *reminds us not to let the perfect be the enemy of the good*

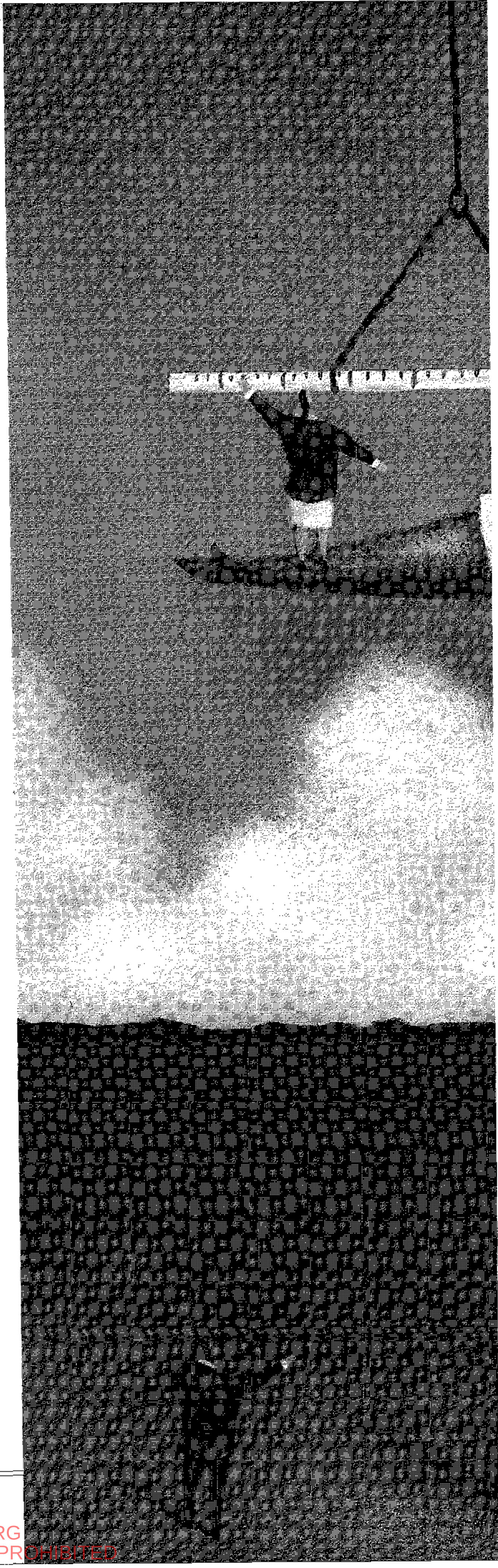
forum

Sizing UP TEST SCORES

by DALE BALLOU

ONE OF THE BASIC CRITIQUES OF USING TEST SCORES for accountability purposes has always been that simple averages, except in rare circumstances, don't tell us much about the quality of a given school or teacher. The high scores of students in a wealthy suburban New Jersey school will reflect the contributions of well-educated parents, a communal emphasis on academic achievement, a stable learning environment at home, and enriching extracurricular opportunities. Likewise, the low scores of students in an inner-city Newark school will reflect the disadvantages of growing up poor. The urban school might have stronger leadership and a more dedicated teaching staff, yet still score substantially lower than the suburban school. As a result, in the past decade researchers have grown interested in ways of measuring and comparing the *gains* in academic achievement that a school or teacher elicits—in other words, a school or teacher's "value added." Say, for instance, that a school lifts its students from the 35th percentile on national tests to the 50th percentile. An accountability system that uses value-added assessment might judge this school more effective than a school whose students consistently score at the 60th percentile. A value-added system

The latest innovation in measuring the performance of schools and teachers holds great promise, but the idea is still way ahead of our ability to execute it



ILLUSTRATIONS BY JOHN BERRY

might also identify a school's best and worst teachers by tracking their students' gains in the course of a year. The prospect of measuring the contribution made by schools and teachers to their students' progress is winning a growing number of converts to value-added assessment. However, some practical complications stand in the way.

It is important, first, to distinguish between assessment for diagnostic purposes and assessment as a mechanism of accountability. Value-added assessment has demonstrated its value in the former capacity. Pioneering work in Dallas and in Tennessee has shown that value-added assessment provides information that can be useful when viewed in context by educators who understand local circumstances.

It is important to distinguish between assessment for diagnostic purposes and assessment AS A MECHANISM OF ACCOUNTABILITY.

The more serious difficulties arise when value-added assessments are used to hold schools and teachers accountable, with high-stakes personnel decisions to follow. The danger is that such assessments will be used to supplant local decisionmaking, rather than to inform it. Unfortunately, our instruments of assessment are not precise or dependable enough for this purpose. I will discuss three problems: 1) current methods of testing don't measure gains very accurately; 2) some of the gains may be attributable to factors other than the quality of a given school or teacher; and 3) we lack a firm basis for comparing gains of students of different levels of ability.

✦ *Measured gains are noisy and unstable.*

Tests are not perfect measures of student ability or achievement. A student's performance on any given test will be due partly to true ability and partly to random influences (distractions during the test, the student's emotional state that day, a fortuitous selection of test items, and so on). This test "error" causes problems enough when we attempt to assess a student's *level* of achievement. The problems are significantly compounded when we take it a step further, to measuring achievement *gains*. A gain score is the difference between two test scores, each of which is subject to measurement error. The measurement errors on the two tests, taken months apart from each other, are unlikely to be related (after all, these are random influences). When we subtract one score from another, the measurement errors do not cancel out. However, a student's true ability does not change that much from one test occasion to another. When we subtract one score from another, a good deal of the portion of the scores that represents true ability *will* cancel out. The result: the proportion of a gain score that represents measurement error is magnified vis-à-vis the initial scores. In statistical parlance, gain scores are much noisier than level scores.

Statisticians facing this problem have adopted procedures that adjust raw measures of gain to minimize the contribution of statistical noise. The noisier the data, the less weight is placed on measured gains for any one school (or teacher). In extreme cases, the school or teacher in question is simply assigned the average level of effectiveness. Of course, the amount of noise in the data is itself something that must be estimated. As a result, these statistical methods are quite sophisticated. Virtually no one who is evaluated by these methods—teachers or administrators—will understand them. Thus, value-added systems that adjust for the unreliability of raw test scores will fail one of the criteria that educators have deemed important for accountability: that they be *transparent*. Measured performance (as determined by the statistical models) will not accord with the raw data. It will be impossible to explain to the satisfaction of educators why two schools (or teachers) with similar achievement gains nonetheless received different ratings of their effectiveness.

Moreover, inequities will arise simply because measured gains are more dependable for schools and teachers for whom there are more data. There will not be enough information about teachers who are new to a school system to obtain reliable estimates of their effectiveness based on past performance—they will simply be deemed "average." Likewise, it will be considerably more difficult for a small school to rank high: it will have to outperform larger schools in order to appear equally effective. (In the same way, a small school's inferior performance may go undetected.) Discrepancies will also arise across subjects. For reasons probably due to the home environment, more of the variation in student reading performance is independent of school quality than is the case in math performance. As a result, it is harder to detect particularly strong (or weak) performance by reading instructors than by math teachers.

In the end, using sophisticated methods of value-added assessment may not be worth the trouble if the object is to identify and reward high performance. William Sanders, formerly of the University of Tennessee and now at the SAS Institute, has done pioneering work to develop a system of value-added assessment, using the results of annual tests administered to all elementary and middle-school students in Tennessee. The great majority of teachers assessed by this system do not differ from the average at conventional levels of statistical significance. A recent investigation of achievement in one large Tennessee school district (in which I am collaborating with Sanders and Paul Wright of the SAS Institute) has found that 20 percent of math teachers are recognizably better or worse than average by a conventional statistical criterion. By the same criterion, the percentage falls to 10 percent in language arts instruction and to about 5 percent among reading teachers. Those who want to reward teachers on the basis of measured performance should consider whether it is worth the trouble and expense to implement value-added assessment if the only outcome is to reward small numbers of teachers. Of course, it is possible to disregard statistical criteria and reward the top 10 percent of teachers in all subjects willy-nilly. But then many rewards will be made on the basis of random fluctuations in the data.

✦ *Gain scores may be influenced by factors other than school quality.* Value-added assessment has one signal merit: it is based on student progress, not on the level of achievement. Schools and teachers are accountable for how much students gain in achievement. They are not given credit for students entering at a high level or penalized when their students start far behind. In effect, value-added assessment "controls for" the influence of family income, ethnicity, and other circumstances on students' initial level of achievement.

However, this may not be enough. The same factors may influence not just the starting level, but also the rate of progress. Thus, even in value-added assessment, it may be necessary to control explicitly for these factors (or demonstrate that they do not matter).

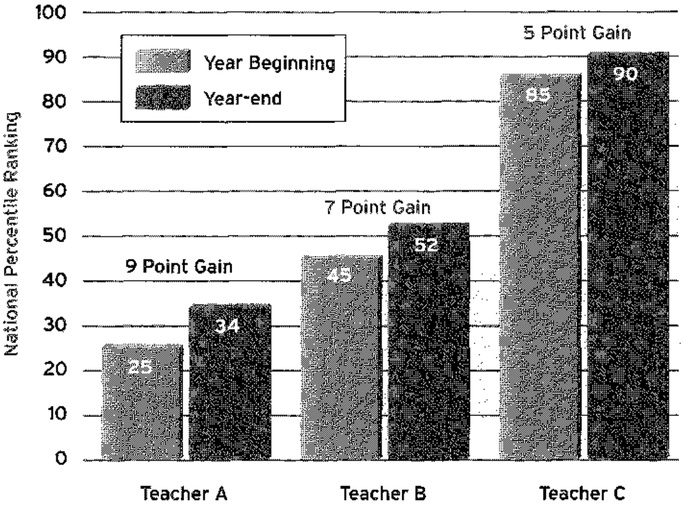
The socioeconomic and demographic factors that might influence student progress make a long list. In practice it is unlikely that an assessment system will have access to data on student backgrounds beyond what is routinely collected by school systems: the percentage of students with limited English proficiency, the percentage eligible for free and reduced-price lunch, and the ethnic and racial composition of the student population. Clearly other factors also matter. Critics of high-stakes assessments will object that without an exhaustive set of controls, the assessment system will end up penalizing some teachers and schools for circumstances beyond their control. Unless it can be shown that value-added assessment need not account for these other influences, schools that receive low marks will have an obvious excuse: the assessment did not recognize that "our students are harder to educate."

Some progress has been made in this task. Sanders, Wright, and I have found that introducing controls for students' race, eligibility for free and reduced-price lunch, and gender (together with the percentage of a teacher's students eligible for free and reduced-price lunch) usually has only a minor impact on teachers' measured effectiveness. However, this study was limited to one school district and one series of achievement tests. Whether the results will generalize remains to be seen.

Moreover, even small differences in measured effectiveness can have practical consequences for schools and teachers, depending on how these assessments are used. For example, suppose it is school policy to reward teachers who score in the top 10 percent. Whether a specific teacher falls into this category can be rather sensitive to the inclusion (or omission) of controls for student background. Even relatively modest changes in measured effectiveness, such as our research has found, can have a decisive influence on whether a teacher falls above or below a cut-off point defined in this manner. A teacher who would rate in the top 10 percent on one measure has only to fall slightly below the cut-off on the other measure to drop out of the category of teachers who are recognized for their excellence. Our findings suggest that this will happen with some fre-

Which Teacher Added the Most Value? (Figure 1)

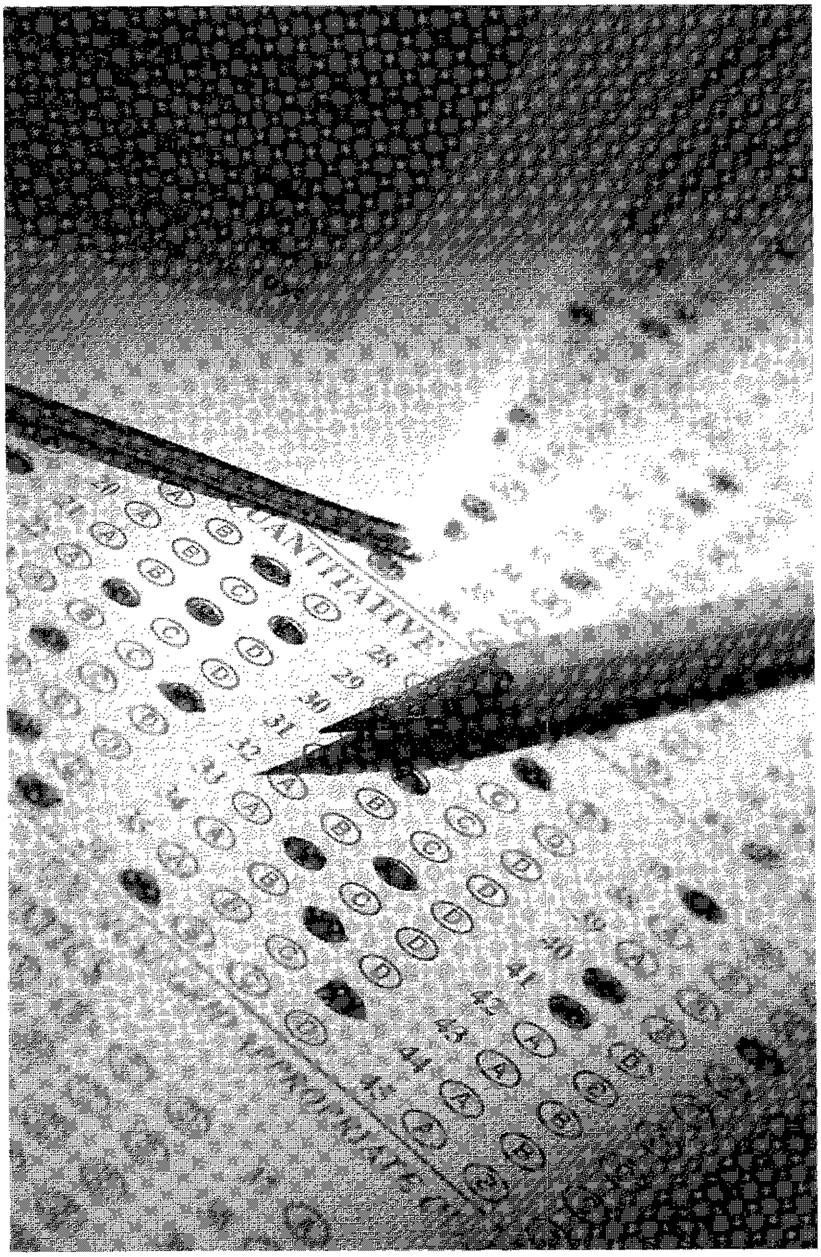
It's unclear whether gains at the top of the scale are truly smaller than gains at the lower end, or vice versa.



Scores at the Beginning and End of the School Year

Those who look to value-added assessment as the solution to the problem of educational accountability are likely to be disappointed. THERE ARE TOO MANY UNCERTAINTIES AND INEQUITIES.

Schools that receive low marks will have an obvious excuse: the assessment did not recognize that "OUR STUDENTS ARE HARDER TO EDUCATE."



quency: more than one-third of the teachers who ranked in the top 10 percent when our assessments included socioeconomic and demographic controls no longer belonged to that category when these controls were omitted from the analysis.

✦ *Similar gain scores are not necessarily comparable.* To practice value-added assessment, we must be able to compare the achievement gains of different students in a meaningful way. We need to be assured that the scale on which we measure achievement is one of equal units: one student's five-point increase on an achievement test, from 15 to 20, must represent the same gain as another student's five-point increase from 25 to 30 (see Figure 1). If it does not, we will end up drawing false conclusions about the relative effectiveness of these students' teachers and schools.

Mathematicians who specialize in measurement in the social sciences, together with experts in the construction and interpretation of tests—psychometricians—have devoted considerable attention to this matter. Their findings are highly unfavorable to value-added assessment. First, it is clear that a simple tally of how many questions a student answered correctly will not have the desired property. Test questions are generally not of equal difficulty. Raising one's score from 15 to 20 might well represent a different achievement gain than an increase from 25 to 30, depending simply on the difficulty of the additional questions that have been answered. This objection also applies to several popular methods of standardizing raw test scores that fail to account sufficiently for differences in test items—methods like recentering and rescaling to convert scores to a bell-shaped curve, or converting to grade-level equivalents by comparing outcomes with the scores of same-grade students in a nationally representative sample.

In the 1950s, psychometricians began to deal with this issue in a systematic way. The result has been the development of "item response theory," which is used to score the best-known and most widely administered achievement tests today, such as the CBT/McGraw-Hill Terra Nova series and the National Assessment of Educational Progress. In item-response theory, the probability that a student will answer a given item correctly is assumed to depend on the student's ability and on the difficulty of the item, as expressed in a mathematical formula. Neither a student's ability nor the difficulty of the item can be directly observed, but both can be inferred from the pattern of answers given by a particular student as well as by other students taking the same test. For example, as one would expect, the more students who answer a given item correctly, the "easier" the item is judged to be. The estimate of a student's ability (known as the *scaled score*) is expressed on the same scale as item difficulty.

The critical question, given that neither ability nor item difficulty can be measured directly, is whether the procedures of inference are powerful enough to put the resulting ratings of ability and difficulty on equal-unit scales. Has student A, whose scaled score rose from 500 to 600 (using item-response theory methods), truly learned less than student B, whose scaled score rose from 300 to 450? Does 100 points at one range of the scale really represent less learning than 150 points at another point on the scale? The fact that we express both scores numerically predisposes us to answer affirmatively. The dubiousness of such a response can be appreciated by approaching the question from another angle, taking advantage of the fact that ability is measured on the same scale as item difficulty. Suppose a student has answered test item A correctly, which has a difficulty rating of 500. Item B is harder, with a difficulty of 600. Clearly the student needs to know more to answer question B than question A. But is the extra knowledge required to answer question B truly less than the extra knowledge required to answer item C, with measured difficulty of 450, compared to item D, with a difficulty of 300? Does a numerical difference of 100 between items A and B really mean that the latter item "contains" 100 more units of something called difficulty—and that these are the same units in which the difference between items C and D is judged to be 150? Do we possess such a scale for difficulty, or are we merely able to determine the order of difficulty, assigning higher numbers to items judged to be harder?

Generally speaking, the latter account of the matter is the correct one. And because ability is measured on the same scale as difficulty, the same holds true of it. In practice, psychometricians usually act as if ability scores are on an equal-unit scale (or, in technical terms, an "interval" scale). But this is merely an assumption of convenience. As prominent psychometricians have pointed out, many of the usual procedures for comparing achievement gains yield meaningless results if the ability scales lack this property.

In the previous example, the size of the intervals, 100 and 150, depended on the choice of the mathematical function expressing the probability of a correct response. Different choices for that function will produce different scales. And choice is just what takes place. There is no "correct" choice—or, more precisely, we possess no criteria for determining whether one choice is more correct than another. Statistical properties, such as the "fit" of the data to the model, are of no help here. As noted in a 1986 article in the *Journal of Educational Measurement* by Wendy Yen, formerly the chief research psychologist with CBT/McGraw-Hill and now with the Educational Testing Service (ETS), there are infinitely many nonlinear transformations of the ability scale that will fit the data equally well, yielding the same probabilities. Yet these transformations will shrink the scale over some ranges and expand it over others, so that student A appears to make more progress than student B using one scale, but less using another.

Our efforts to determine which students gain more than others turn out to depend on arbitrary choices that MAKE SOME EDUCATORS LOOK BETTER THAN OTHERS.

These conclusions cut the ground out from under value-added assessment. Our efforts to determine which students gain more than others—and thus which teachers and schools are more effective—turn out to depend on conventions (arbitrary choices) that make some educators look better than others. This does not mean that testing is of no value. It is still possible to rank students' performance and to ask how many students exceed a specified benchmark. But the finer kinds of measurement required to compare the progress of students at different levels of initial ability exceed the capacities of our instruments. As Henry Braun of ETS wrote in a 1988 article for the *Journal of Educational Measurement*: "We have to be very careful about the questions we ask and very sensitive to the possibilities for obtaining misleading answers to those questions. . . . In particular, we should probably give up trying to compare gains at different places on the scale for a given population."

Conclusion

There is a simple idea behind value-added assessment: schools and teachers should be evaluated based on student progress. As the foregoing discussion shows, however, successful implementation of this concept is far from simple. It is much harder to measure achievement gains than is commonly supposed.

Notwithstanding these problems, policymakers, educators, and the public will continue to look at indicators of student progress to see how schools are doing. For some purposes, this seems entirely reasonable. Sanders's assessment system has been a beneficial diagnostic tool in Tennessee. But those who look to value-added assessment as the solution to the problem of educational accountability are likely to be disappointed. There are too many uncertainties and inequities to rely on such measures for high-stakes personnel decisions.

–Dale Ballou is an associate professor of economics at the University of Massachusetts at Amherst.

Expert MEASURES

by ANITA A. SUMMERS

IN ONE SUBURBAN SCHOOL DISTRICT, TEACHERS ACROSS the system were ranked and evaluated according to the contribution they had made to student learning—based on a value-added analysis of state test results. When they were ranked again the next year, the results were very similar except for one teacher who moved from a top rank to a very low rank. When the school superintendent looked at the results, he immediately identified the teacher—her husband had died during the second year of rankings.

What does this anecdote illustrate? For starters, that value-added assessments can misidentify good and bad teachers—and that the discrepancies can be cleared up by local administrators. More important, however, was the fact that the findings were robust; teachers' rankings were similar across three years of analysis. Moreover, these rankings were used to grant a standard reward to a top group of teachers, not to make fine distinctions in the amount of compensation that teachers received. In other words, the results of value-added analysis, when examined over a period of two or three years, were stable and were used in a way that respected the margin of error involved in using these statistical techniques.

Critics of value-added assessment tend to embrace the concept but don't want the results gleaned from such analysis to be used for accountability purposes—and especially don't want to use the results to reward or sanction teachers. But teachers are the dominant school input, in terms of both spending and impact on student learning. Excluding them essentially leaves the education system without accountability.

The main concern with value-added assessment is that the technique exacerbates the amount of random error involved in measuring student performance. The risk is that teachers and schools may be

All the evidence to date shows that value-added techniques are being employed responsibly

wrongfully rewarded or punished because value-added techniques either over- or underestimated their students' learning gains. However, no intelligent users of standardized testing would make policy choices based on a single year's result or small differences among schools and teachers. Texas, for example, rewards schools based on value-added achievement gains calculated using two-stage regression analysis. The state does not hand out awards based on decimal-point differences among schools; officials reward a previously set percentage of top-ranking schools. The point here is that most of the statistical objections to value-added measurements assume a misuse of the analysis. The statistical "noise" involved in measuring value-added should preclude decisions that are based on small, unreplicated analysis; it should not preclude decisions that are based on gross findings.

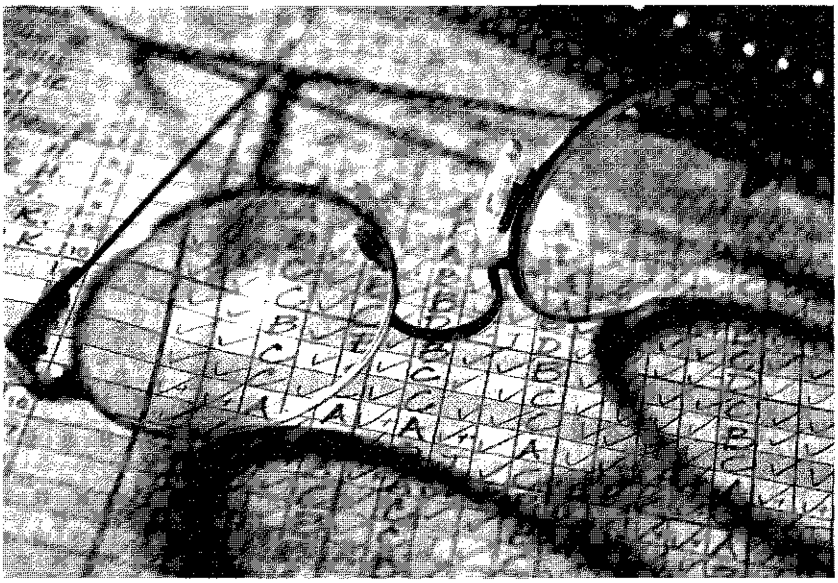
There is already considerable evidence that low-achieving students are the main beneficiaries of the CHANGES THAT OCCUR WHEN VALUE-ADDED ASSESSMENT IS USED.

Confidence in gross findings can be developed by replication, by averaging results over several time periods, and by using several measures of the development of human capital—not tests alone, but also attendance rates, dropout rates, and promotion rates (a very high-quality assessment will track indicators of human capital such as post–secondary school earnings and higher-education outcomes as well). The richer the measures used, the less weight there is on the psychometric concerns involving test scores. The alternative is to rest teacher compensation on factors that have little to do with student learning. It is now well established, for example, that the number of degrees teachers possess and the number of hours teachers spend in education courses are unrelated to student learning. Put another way, an important criterion in the determination of a teacher's salary does not have any bearing on the ability of the teacher to develop human capital. We know that because it has been replicated in many studies in many school districts, even though salary schedules have yet to reflect this information.

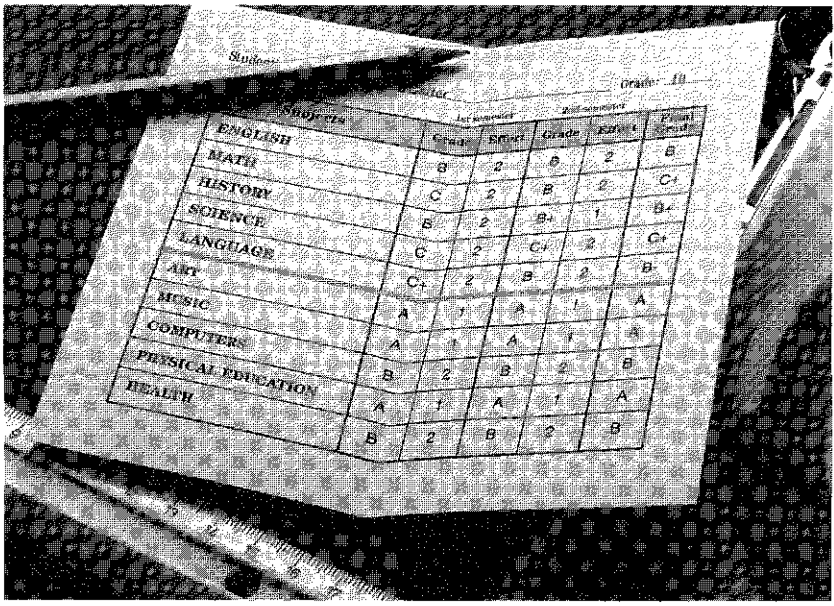
Critics often cite the difficulty of comparing the results of large and small schools and comparing one subject with another. It is essential in the debate over the usefulness of the value-added assessment approach that the unit of observation and the comparison group be specified. If, as has been the case in a number of places, the comparison group is all the teachers in a given grade in the school district—with, say, the top 15 percent of the 4th grade teachers receiving an award—what is the significance of a big or small school? The class is the unit, and class size tends to be uniform within a school district. At the high-school level, the comparison group is likely to be, for example, all history teachers or all science teachers across the district. The award-receiving group is a percentage of that comparison group, and it is not affected by the test scores in another subject.

If multiple factors are used for assessment, efforts to check on robustness are made, only extreme performances are rewarded or sanctioned, and comparison groups are selected carefully, the technical psychometric points raised by critics should not swamp the incentive and information benefits of performance-based compensation plans for teachers. If we are ready to determine reading programs, language labs, class sizes, and the use of computerized learning on the basis of value-added assessments, we should be ready to reward teachers using the same techniques.

There is already considerable evidence from several places—such as Tennessee and Florida, where value-added analysis has been used for accountability purposes—that low-achieving students are the main beneficiaries of the changes that occur when these techniques are implemented. When low-achieving students are taught the same body of knowledge over and over again, and when they are taught how to work under a time constraint, they benefit. Value-



added assessment techniques reveal that information.



Student

Grade: 10

Subjects	1st semester Grade	1st semester Effort	2nd semester Grade	2nd semester Effort	Final Grade
ENGLISH	B	2	B	2	B
MATH	C	2	B	2	C+
HISTORY	B	2	B+	1	B+
SCIENCE	C	2	C+	2	C+
LANGUAGE	C+	2	B	2	B-
ART	A	1	A	1	A
MUSIC	A	1	A	1	A
COMPUTERS	B	2	B	2	B
PHYSICAL EDUCATION	A	1	A	1	A
HEALTH	B	2	B	2	B

The Illusion of Transparency

Critics of value-added assessment say that it is just too complicated for teachers and the public to understand the results these systems will generate. In other words, the results will not be transparent. But if transparency is the criterion that trumps all other criteria, we would be compelled to use an assessment method that we know to be wrong. For example, there is no satisfactory way to make judgments about which method of teaching reading is superior—whole language or phonics—without factoring in the socioeconomic, school, and teacher characteristics of each of the groups of students in the experiment. Statistical controls must be used if the assessments of teachers, schools, or programs are to be accurate, even though very few educators understand the statistical principles and methods involved. I do not require a transparent understanding of the efficacy of the flu shot I take, nor do I require a transparent understanding of the operating characteristics of my car; I trust the experts on the techniques. So must it be in educational evaluation.

I do not require a transparent understanding of the operating characteristics of my car; I trust the experts on the techniques. SO MUST IT BE IN EDUCATIONAL EVALUATION.

The problems with the use of value-added assessments, even for teachers, are greatly exaggerated, and the alternatives are simply untenable.

One proposed alternative is the use of one or more methods of subjective evaluation. Other teachers, students, and/or parents can be surveyed to make the judgments. Most of those surveys focus on whether respondents like the teacher, are happy in the classroom, or equally soft attributes. They do not really determine what kind of learning gains a certain teacher is eliciting relative to other teachers. Most people look back at their primary and secondary schooling and identify an extremely demanding teacher (whom they disliked at the time) as the one who made the biggest contribution to their educational development.

A second possibility is to set up a standard—some threshold—of student achievement as an absolute hurdle to define adequate performance. The problems with this are legion. How should the threshold be determined? Those who want to be rewarded set lower thresholds than those who watch budgets. The threshold becomes a major bargaining tool, quite divorced from the informed decisionmaking objective. What incentives for improvement are there for those who crossed the threshold?

Basically, subjective evaluation allows the information essential to the rational allocation of educational resources to be derived politically rather than scientifically.

The state of American public education has been deplored by critics for many years. The notion that student learning has remained somewhat stagnant over the past century, though the country has spent steadily more real resources on education, is a matter of profound concern. This surely indicates that we do not know a lot about what works and what doesn't.

We need to know if we are to change the pattern. We must have calibrated results; we must use sophisticated statistical methods to interpret the data; we need to use multiple measures of performance; and we need to implement the analysis in ways that are appropriate to the quality of the information. And teachers, the most important school-controlled input into the educational process, cannot be exempt from this information-gathering activity. Our current unhappy results are consistent with the subjective and/or unsophisticated tools we use for assessing effectiveness and creating incentives.

–Anita A. Summers is a professor emeritus of public policy and management at the Wharton School at the University of Pennsylvania.

THE Business Model

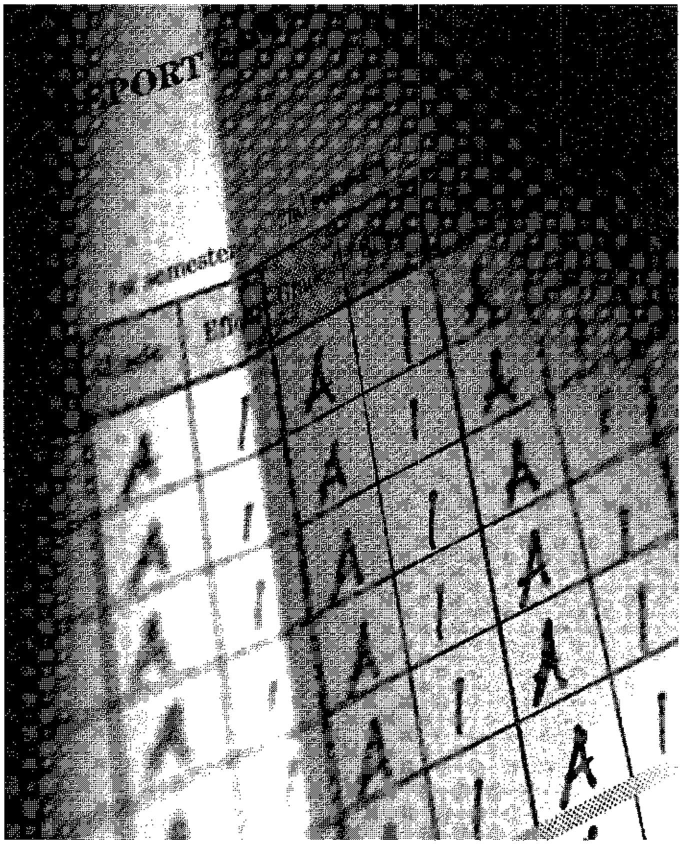
Value-added analysis is a crucial tool in the accountability toolbox—despite its flaws

by JAY P. GREENE

LIKE THE MAKERS OF HOT DOGS, PSYCHOMETRICIANS, economists, and other testing experts know too well what goes into the creation of achievement tests. Their intimate knowledge of the technical difficulties involved in measuring student achievement makes a number of these testing experts some of the most vocal (and persuasive) opponents of testing. But the flaws in techniques like value-added assessment do not automatically lead to the conclusion that those techniques shouldn't be used to hold educators accountable. Testing may be imperfect, but the alternative—the old system, which allowed us to know very little about the performance of educators—is far, far worse.

To be sure, many of the technical criticisms of value-added testing are correct. It's true that there is more random error in measuring gains in test scores than in measuring the level of test scores. It's true that there is some uncertainty as to whether gains in one area of the test scale are equal to gains at another point in the test scale. And it's true that factors besides the quality of the schools can influence the gains that students achieve. But, on balance, these downsides hardly outweigh the benefits to be reaped from being able to measure and reward productivity in education.

Consider what is likely to continue to happen in education without high-stakes value-added assessment. Unless productivity is measured, however imperfectly, it is not possible to reward teachers, administrators, and schools that contribute most to student learning. If we do not reward productivity, we are unlikely to encourage it. If we do not encourage it, we should not expect more of it.



In fact, this is precisely what has been happening in U.S. education during the past few decades. Between 1961 and 2000, spending on education tripled after adjusting for inflation, from $2,360 to $7,086 per pupil. During that time, student performance, as measured by scores on the National Assessment of Educational Progress (NAEP) and high-school graduation rates, has remained basically unchanged. Whenever spending triples without any significant improvement in outcomes, there is a serious productivity crisis. Yet U.S. public schools just keep chugging along, resisting serious attempts at reform.

Meanwhile, private firms in the United States have been able to achieve steady gains in productivity because the discipline of competition has forced them to adopt systems for measuring and rewarding productivity. Firms that fail to measure and reward productivity lose out to their competitors who do.

Moreover, the systems that private companies use to measure and reward productivity are far from flawless. In fact, the challenge of measuring productivity in the private sector is often as great as or greater than in education. Imagine a soft-drink company that wishes to measure and reward the productivity of its sales force. The company might determine bonuses (and even decisions on layoffs) based on its salespeople's success at increasing soda sales in their sales area. Like measuring gains in test scores, measuring increases in soda sales is fraught with potential error. Changes in soda sales could be influenced by a variety of factors other than the sales acumen of an employee. Unusually cold weather in an area, a local economic downturn, or exceptional promotional efforts by competitors could all suppress the soda sales of even a very good salesperson. If data on sales are collected using survey techniques, there is also the possibility of random error attributable to the survey method, just as testing has random error. Moreover, if we are comparing sales increases across geographic areas, it is unclear whether it takes more skill to sell soda in an area where the market is already saturated than in an area that initially consumes less soda.

In short, many of the same technical flaws that critics find in value-added testing also exist in the measurement of increases in soda sales. Changes in outcomes may be attributable to factors other than the efforts of the employee. There is random error in collecting the data. And the effort required to produce gains at one level may not be the same as at another level. The only difference is that private firms have rightly not let their inability to achieve the best deter them from pursuing the good.

In the private sector, companies have realized that even flawed evaluation systems nevertheless encourage improvements in productivity. This is because employees cannot be sure that a flawed system will completely obscure the picture of how hard they're working. Employees therefore act as if their productivity were being measured accurately; the chance that slacking will be detected inspires employees to avoid slacking. In fact, evaluation systems with a fairly large amount of error in measuring productivity can still be effective at motivating improvement—if the errors are mostly random, or at the very least do not create perverse incentives, such as encouraging teachers to focus on improving the achievement of one group of students to the exclusion of others.

None of the technical concerns with value-added testing involve perverse incentives. For the most part, the criticisms have to do with random noise in measuring gain scores. Even the nonrandom errors that worry testing critics, such as unevenness in the testing scale or the possible influence of factors outside the school's control, do not create perverse incentives because there are no strong theories about the kinds of behaviors those errors would encourage.

If no one knows what is being mistakenly rewarded, no one has an incentive to engage in that perverse behavior. As long as educators are aware of what the value-added system is supposed to be rewarding, and as long as that system rewards the desired outcomes more than it erroneously rewards something else, the system will help to elicit more of the desired outcomes—namely, improvements in student achievement.

The Uses of Data

The development of even an imperfect value-added testing system would revolutionize the systems for hiring, promoting, and compensating teachers. Our current methods provide teachers with little incentive to improve achievement. Promotions and salary increases are based on teachers' seniority and their acquisition of advanced degrees. These characteristics are, at best, weakly related to student achievement. Excellent teachers who possess a master's degree and a few years' experience receive exactly the same salary as lousy teachers with the same formal credentials. Under the current system, we have turned the keys over to educators and trusted that their professionalism will yield improvements in student achievement. Education's productivity crisis in the past four decades should be evidence enough that simple trust is not sufficient.

The development of value-added assessment would also revolutionize how we govern schools and hold them accountable. We currently have little rational basis for saying that a particular school is a good school or that a particular superintendent is a good superintendent. Value-added testing would at least give voters some idea of whether they are getting their tax money's worth out of the school system by giving them at least some information on how the schools are doing. The fact that voters would have better information on achievement provides the school board with incentives to hire and retain a superintendent who can elicit improvement in student learning. The superintendent, in turn, has an incentive to hire and retain principals who will use the value-added results to hire and promote the best teachers.

In the private sector, companies have realized that even flawed evaluation systems nevertheless ENCOURAGE IMPROVEMENTS IN PRODUCTIVITY.

Critics of value-added assessment don't necessarily object to using value-added assessment. They object to using the data gleaned from it for high-stakes purposes, such as rewarding or punishing individual schools and teachers. Instead, they suggest that value-added results be provided to administrators so that they can make informed decisions about their employees. This is, to some extent, what happens in the private sector; most private firms do not use the crude techniques exemplified by the real-estate company in David Mamet's *Glengarry Glen Ross*, where the salesperson with the fewest sales was fired. Most companies use productivity measures to inform the subjective assessments of supervisors, with some companies permitting less subjective judgment than others for fear of bias or favoritism.

But here is where the parallel between measuring productivity in public education and private industry ends. Supervisors in private companies have incentives to use the information provided by productivity measures properly, because their companies face the discipline of competition from other companies. If supervisors fail to put data on costs, sales, and revenue to good use, their companies will lose out to competitors who do.

In public education, by contrast, local decisionmakers have few or no incentives to make good use of data in assessing their employees because public schools face no meaningful competition. There are basically no consequences for principals who disregard the results of value-added assessments in making decisions about employees, and they're more likely to disregard those results if they consider value-added assessment an unreliable analytical technique. Superintendents will not be able to judge whether principals have used their discretion properly, because they will be told that the value-added test results are not proper grounds for assessing the decisions of principals. And school boards and voters, in turn, will all be stymied in making independent judgments because they will be told that the professional decisions of educators are more reliable than value-added test results.

Under the current system, excellent, hard-working teachers who put in tons of overtime receive the same salary as mediocre teachers. WHAT'S FAIR ABOUT THAT?

So there is good reason to fear that principals in public schools, if given discretion to reward teachers as they please, will base their decisions on personal relationships rather than on the results of value-added assessments. In both private industry and public education, a balance must be struck between the mechanical use of productivity measures in assessing employees and relying on the subjective judgments of supervisors. But, in public education, the balance needs to tilt more toward the mechanical application of results, because supervisors have fewer incentives to make appropriate subjective judgments. This means that value-added assessment needs to be high stakes to have the desired positive effect on student learning.

Injustices are unavoidable if value-added assessments are used more mechanically. Some educators will be improperly punished for eliciting what appear to be low gains because of measurement error. Conversely, some educators will be rewarded for improvements for which they were not actually responsible. That said, some of the flaws in value-added assessment have potential technical solutions. For example, if judging a teacher based on his classroom's test scores contains too much error because the sample is too small, we might decide to rate teachers based on a moving average of multiple years of results, thereby increasing the sample size and reducing the random error. But even with technical fixes, some injustices will still occur.

This is not a good reason to abandon the idea, however. After all, some educators will be treated unjustly under any evaluation system. Under the current system, excellent, hard-working teachers who put in tons of overtime receive the same salary as mediocre teachers. What's fair about that? The Buffalo, New York, district recently announced layoffs as a result of a budget deficit. Who were the first to receive pink slips? Not the worst teachers in the district, but the most recently hired. Surely some excellent teachers lost their jobs, while the district retained its burned-out veterans. A peer- or supervisor-review system may reward teachers who are popular among their peers rather than effective with students. Attempting to measure and reward successful educators, with all of its imperfections, is likely to create fewer injustices than any other arrangement. At least using value-added assessments increases the chance that good teachers are rewarded and bad teachers are sanctioned.

Besides, ensuring that every single educator receives justice is at most a secondary concern. An obvious but infrequently recognized truth is that the primary purpose of the education system is to provide a quality education to all students. If high-stakes value-added assessment can help motivate public schools to provide students with a better education, it's a promising reform even if it has some cost to school employees. In no other industry would we even entertain the notion that the interests of employees trump those of customers. Only the political dominance of teacher unions makes us consider the question. It is true that customers usually receive the best service when employees are treated well and fairly. Happily, high-stakes value-added assessment is likely to achieve both ends.

The productivity crisis in public education has certainly created injustices for students and taxpayers during the past few decades. Students have failed to receive high-quality instruction while taxpayers have been paying more and more in return for stagnant test scores. Again, what's fair about that?

–Jay P. Greene is a senior fellow at the Manhattan Institute for Policy Research.

Enemy OF THE GOOD

No standardized test is perfect. But they're useful nonetheless

by DONALD R. McADAMS

VALUE-ADDED ASSESSMENT IS FLAWED, EVEN SERIOUSLY flawed. Nevertheless, I remain committed to the use of testing and value-added techniques to improve America's public schools. Voters and policymakers should hold school districts and their employees accountable for student learning. Value-added high-stakes tests, with all their flaws, are an excellent way to do this.

Of course, tests don't stand alone. To have value they must be aligned with content standards, performance standards, and accountability. And principals and teachers must have the freedom and flexibility to meet those standards, while school districts provide them with adequate resources and support.

High-performance organizations measure almost everything. Why? Because measuring changes behavior. The best way to focus the attention of a workforce on something important is to measure it. As critics points out, measuring the performance of schools and teachers is difficult, very difficult. Does this mean we shouldn't do it? Does this mean we leave teachers, in the privacy of their classrooms, to set their own standards, develop their own performance measures, and tell us whether the children are learning? We tried this. The result: too many teachers neglected to teach the curriculum or did not teach effectively, and too many children suffered the consequences.

Yes, we should recognize that a test measures true ability *and* random influences. Yes, measured gains are noisy and unstable. Yes, socioeconomic and demographic factors have some influence on measured student progress. And indeed, scaling is a problem. No standardized test is perfect, yet we use them all the time, and to good effect. Physicians, lawyers, accountants, financial planners, real-estate brokers, and pilots all take high-stakes tests. These tests ensure that professionals have the knowledge necessary to serve the public well.

Teachers know that standardized tests are not perfect measures of what their students have learned, just as they know that the assessments they develop for their own use are not perfect measures. Yet they still use them to diagnose, motivate, and focus classroom learning. And how often are they surprised by a child's standardized test score? Usually it is just about where they expected it would be.

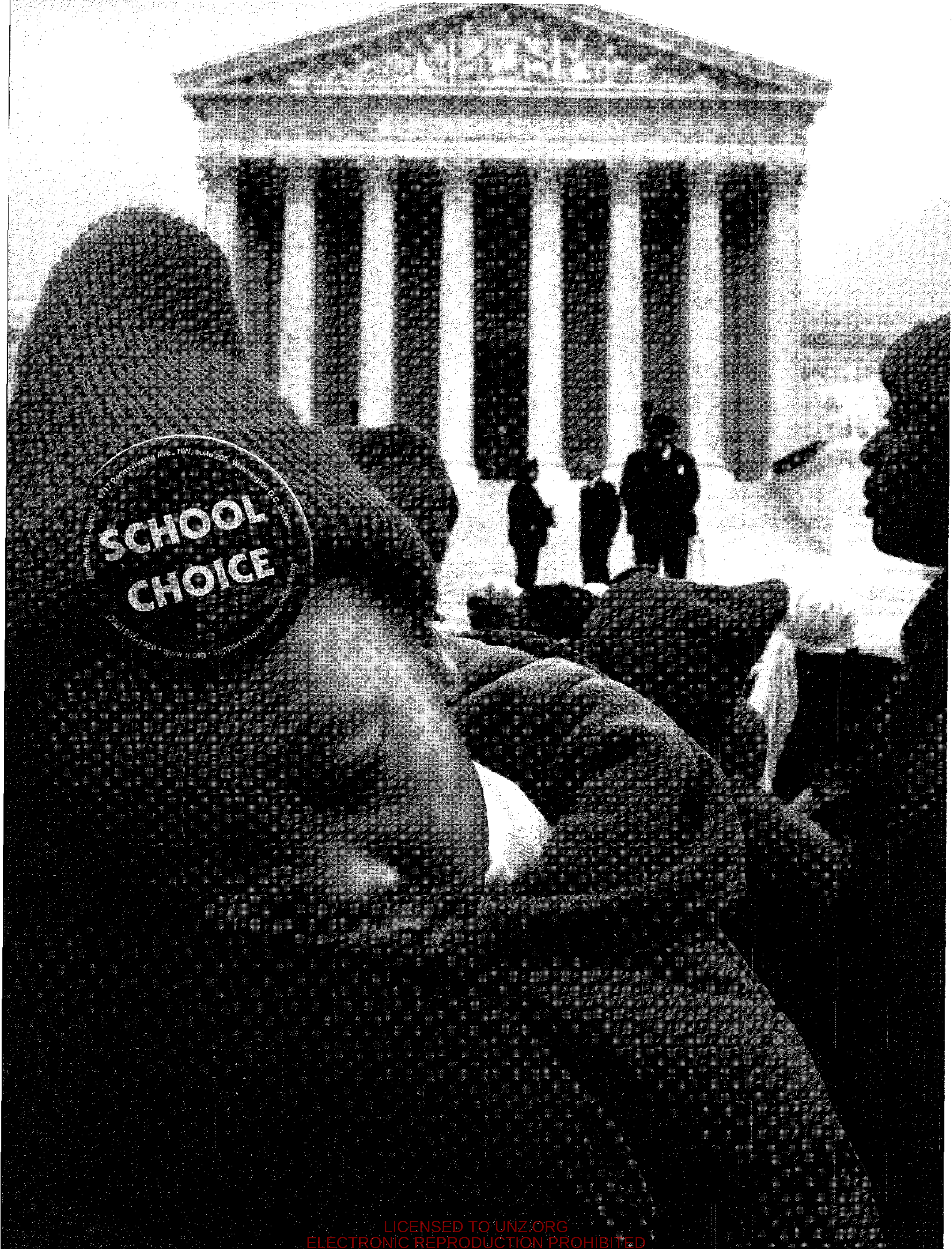
Standardized high-stakes tests also don't measure school improvement perfectly, and they shouldn't be the only accountability device we use. Nor should they be the sole measure of teacher effectiveness. But imperfect as they are, standardized tests do the job. They enable policymakers and the public to answer much more confidently the question, "Are the children learning?" More important, they change behavior.

The constant measurement of student achievement focuses everyone's attention on student achievement. Superintendents, principals, and teachers now spend more time trying to link the structure and work of the organization to student learning. Discipline creeps back into the organization. Practices that don't seem to improve student achievement are dropped, and practices that do work spread throughout the organization. Innovation begins to flourish. Student achievement improves.

As a former 12-year school board member for the Houston Independent School District, I have seen this happen in Houston and in Texas. Demanding accountability for results and measuring achievement with the Texas Assessment of Academic Skills (TAAS), a criterion-referenced assessment—actually, a rather blunt instrument—has spurred significant improvement in student achievement. This improvement has been displayed not only on the TAAS; it has shown up in Houston on the Stanford 9 and statewide on the National Assessment of Educational Progress.

Obviously, test data can be given more weight than they deserve, and there is a danger that administrators might use test data inappropriately to make personnel decisions. But there seems to be little evidence that this has happened and a huge body of evidence that high-stakes tests of all kinds focus school systems on teaching and learning. So let the psychometricians continue to improve techniques for value-added assessment. We need them to do so. And let economists and statisticians continue to point out the flaws in these tests, so that policymakers don't misuse them. Meanwhile, policymakers should deepen and broaden their commitment to standards-based reform and high-stakes tests. The benefits greatly outweigh the risks.

–Donald R. McAdams is executive director of the Houston-based Center for Reform of School Systems and the author of Fighting to Save Our Urban Schools . . . and Winning! Lessons from Houston *(Teachers College Press, 2000).*


SCHOOL
CHOICE

Will the U.S. Supreme Court's decision in Zelman *end the voucher debate?*

VOUCHERS on TRIAL

by
JOSEPH P. VITERITTI

LEGAL EXPERTS ARE ALREADY DRAWING ANALOGIES between *Zelman v. Simmons-Harris* and landmark rulings like *Brown v. Board of Education* and the seminal First Amendment decisions that have shaped American jurisprudence over the past half century. How the U.S. Supreme Court rules on the constitutionality of Cleveland's school-voucher program will not only define the legal boundary between church and state more clearly; it could also help redefine the meaning of public education and expand the range of opportunities available to poor children.

Enacted by the Ohio legislature in 1995, the Cleveland Scholarship and Tutoring Program allows 4,000 low-income children to attend private religious and secular schools with up to $2,250 in public support. Participating schools must cap their tuition at $2,500 a year; the state pays up to 90 percent of whatever the school charges, depending on family income. Following a high-profile legal battle, the program was upheld by the Ohio Supreme Court in 1999, prompting opponents to take their case into federal court. On the day before school was to open that year, federal district court judge Solomon Oliver struck down the program, ruling that the use of tax dollars to pay for children to attend religious schools offends the First Amendment's Establishment Clause. Judge Oliver halted the acceptance of new students to the program while the case was being appealed. Soon thereafter, a sharply divided (5–4) U.S. Supreme Court took an unusual move to vacate Judge Oliver's injunction and

SCHOOL
CHOICE
NO

PHOTOGRAPHS BY STEPHANIE KUYKENDAL

STRENGTHEN OUR PUBLIC SCHOOLS
EVERY CHILD

Challengers pointed out that most of the schools involved in the voucher program (46 of 56, accounting for 96 percent of the students) were religious.

allow the program to continue unaltered while the case is in litigation.

In December 2000, the U.S. Court of Appeals for the 6th Circuit affirmed Judge Oliver's ruling by a 2–1 vote. The appeals court relied heavily on legal precedents set down by the Supreme Court in 1973 in *Committee for Public Education v. Nyquist*. At issue in *Nyquist* was a New York State program that gave low-income parents a partial tuition reimbursement for private-school tuition. The *Nyquist* Court found that the tuition-grant program had the "impermissible effect of advancing religion." It concluded that direct or indirect aid to sectarian schools is essentially a government-subsidized incentive to practice religion.

The thinking in *Nyquist* was remarkable on several counts. Inherent in the incentive concept is the assumption that parochial schools are so superior to public schools that the opportunity to attend the former is irresistible, even to those parents who do not want their children educated in a religious environment. Reasonable people can conclude that the lure of a safe and sound education is an argument for choice rather than against it. Preoccupied with the religious character of parochial schools, the majority also presumed that the court is capable of looking into the minds of legislators to determine their motivations. Using effect to derive intent, the court concluded that incidental aid to religious institutions in the form of tuition relief to parents is tantamount to a purposeful government act to promote religion.

There were specific facts pertinent to the *Zelman* case that the appellate panel deemed relevant to the incentive argument. Challengers pointed out that most of the schools involved in the voucher program (46 of 56, accounting for 96 percent of the students) were religious, leaving few secular options available for participating families. The appeals court accepted this argument even though the Supreme Court had acknowledged similar circumstances in 1983 (*Mueller v. Allen*) when it upheld a Minnesota program that gave a tax deduction to parents for tuition and other education expenses. While recognizing that most of the deductions were used for parochial school tuition, the *Mueller* Court found that because parents could deduct expenses for public, private, or religious schools, the deduction was neutral toward religion.

In *Zelman*, the Ohio attorney general further pointed out that schools participating in the Cleveland voucher program represent only a small portion of the range of choices available outside the regular public schools. In 1999 Cleveland had 23 magnet schools with 13,000 students in attendance and eight charter schools with 1,600 students in attendance, compared with the 3,800 in the voucher program. The two-person majority refused to accept the range-of-choice argument, however, because the magnet and the charter programs were not enacted under the auspices of the voucher law that was being reviewed. Legally speaking, these other choices did not exist. Under the rules of evidence defined by these judges, the same court that could peek into the minds of legislators to determine intent could not recognize hard evidence crucial to determining whether dissatisfied parents at regular public schools had choices beyond religious schools.

Both sides introduced evidence concerning the amount of the voucher. The state of Ohio, in an effort to rebut the incentive argument, explained that the amount ($2,250 maximum) was small in comparison with the per-pupil spending in regular public ($7,746), magnet ($7,746), and charter schools ($4,518). Looking at things strictly from a resource perspective, parents had a disincentive to send their children to schools participating in the voucher program. Opponents argued that since parochial schools were the only nonpublic schools with tuition rates low enough to be covered by the voucher, the program was indeed an incentive to attend these schools. To the extent that the latter argument has merit, the remedy seems obvious: amend the voucher law to make the amount higher, let's say equal to the per capita amount spent in regular public schools. This would have to be done by the Ohio legislature.

There would be a delicious irony to such a resolution. It would certainly appear equitable, in light of the tortured history of school-finance litigation in Ohio. Carried to its logical conclusion, such a resolution might also add charter schools to the mix of institutions eligible for equal funding. Opponents of choice who raise the funding issue as a means of striking down vouchers would not welcome such a remedy, however. It is not financial equity they seek, but the defeat of the voucher law. And school-finance reformers who have spent oodles of time and money in litigation are not likely to receive such a remedy kindly either. For the most part, their sense of fairness applies only to children in public schools.

What Will the Court Do?

Of course, predicting what the Supreme Court is going to do in a particular case can be more difficult than calling the World Series in the middle of May. We begin with certain general expectations based on past performance, all the while knowing that anything is possible. At least baseball has winners and losers; legal contests are more complicated. To say that the Supreme Court will rule one way or another oversimplifies the process of judicial decisionmaking. Crafted to accommodate the philosophies and styles of the individual justices needed to assemble a majority, legal opinions are written with great nuance. Their outcome depends on what question or questions the judges agree to address, and with what level of specificity. As a rule, the Court tends toward more narrow rulings, with deference paid to precedent. But precedents are rarely consistent; and First Amendment case law is among the most inconsonant.

The First Amendment jurisprudence that has unfolded over the past two decades, however, seems to favor the program in question. In the aforementioned *Mueller* case, the Court approved a tuition–tax deduction program in Minnesota.

Some black leaders see school choice as a civil-rights issue—indeed, as a fulfillment of the promise articulated in *Brown v. Board of Education:* to make education available to all "on equal terms."

More recently, the Rehnquist Court has overturned longstanding precedents in order to allow public school teachers to provide remedial services to children on the premises of religious schools (*Agostini v. Felton*, 1997) and parochial schools to receive direct aid in the form of computers and other instructional equipment (*Mitchell v. Helms*, 2000). In 1998, it refused to hear a challenge to a voucher program in Milwaukee that was approved by the Wisconsin Supreme Court.

Still, there is a distinct, though improbable, possibility that the Supreme Court will rule against the voucher program on fundamental First Amendment grounds. This is improbable because since 1986 (*Witters v. Washington*) the Court has adopted

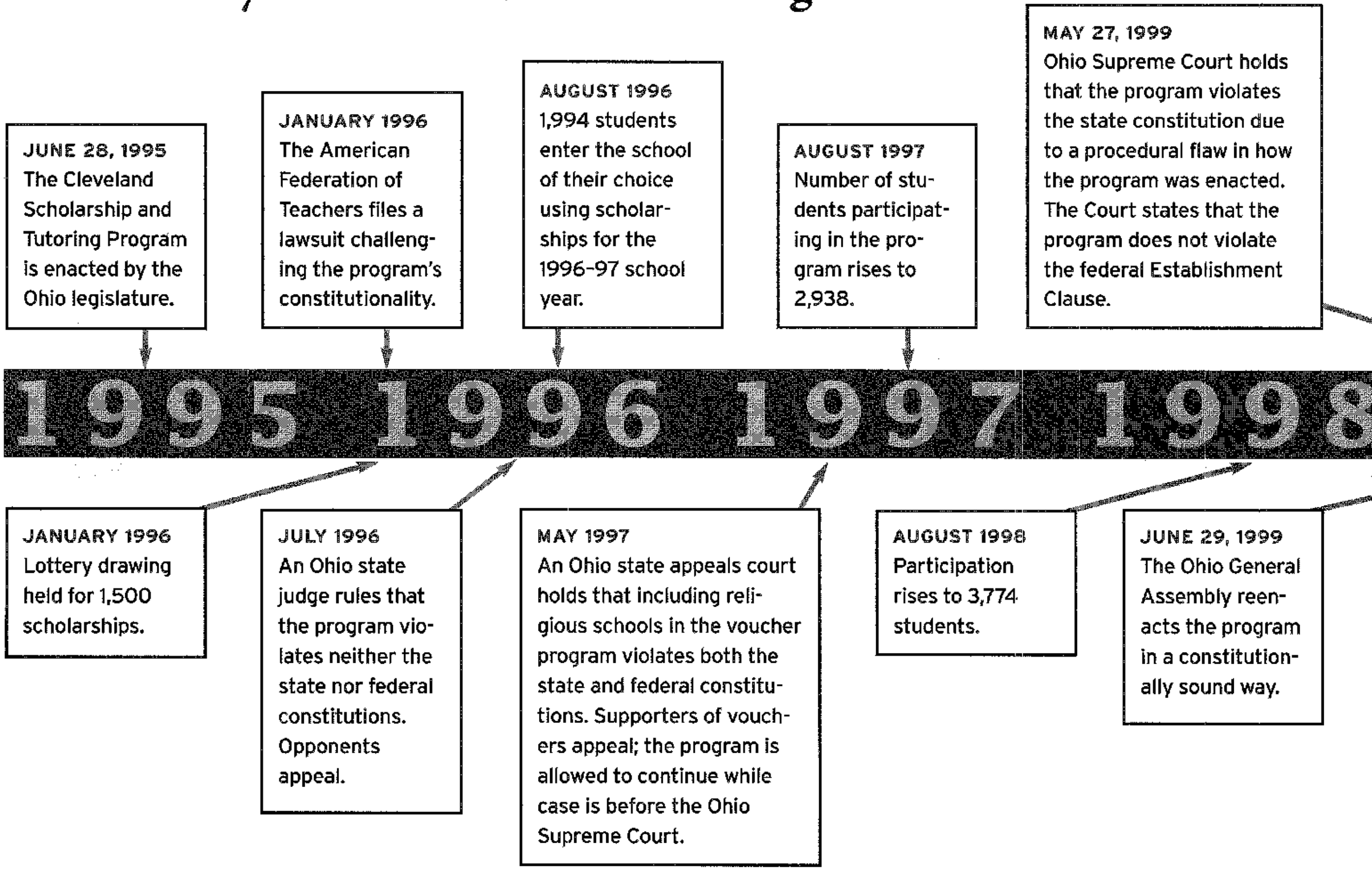
to ME and MY PARENTS
Choice
append
PARENTS CHOICE

guidelines that allow indirect aid to parochial schools as long as the aid is appropriated neutrally and results from independent decisions by parents who select those schools. These more permissive guidelines were drafted in a concurring opinion by Justice Lewis F. Powell Jr., the author of *Nyquist*. The original plaintiffs in *Zelman* argued that parental independence is compromised by an administrative process that sends the voucher check directly to a religious school to be signed by the parent. Again the solution to this problem, if it is really a problem, is rather easy: just send the check to the parent, and let the parent pay the school. The existing procedure was implemented for the sake of administrative expediency. Either way, the money reaches the school because a parent chose that school for her child. But in some courts procedure trumps principle. Legal reasoning would require that the voucher program, operating as it does, be struck down.

Let's imagine, for a moment, what would happen if it were. In Ohio the issue would be turned over to the state legislature, and once again political irony would be the order of the day. The same people who raised procedure as a point of contention in court would do everything possible to preserve the procedure in the law in order to maintain that the program is void. If choice supporters succeeded in changing the procedure, the issue of constitutionality would be resolved. Realistically speaking, however, the fate of the Cleveland program is unlikely to turn on the payment question. There is a bigger question.

While unlikely, it is conceivable that a majority of the justices hearing *Zelman* could agree with the argument that providing unrestricted aid to children attending sectarian schools allows the state to endorse, subsidize, and advance religion. The impact of this ruling would be substantial. Not only would it terminate the voucher program for 4,000 children in Cleveland; it would open to challenge the Milwaukee program through which 10,000

Brief History of Cleveland's Voucher Program



JUNE 28, 1995
The Cleveland Scholarship and Tutoring Program is enacted by the Ohio legislature.

JANUARY 1996
Lottery drawing held for 1,500 scholarships.

JANUARY 1996
The American Federation of Teachers files a lawsuit challenging the program's constitutionality.

JULY 1996
An Ohio state judge rules that the program violates neither the state nor federal constitutions. Opponents appeal.

AUGUST 1996
1,994 students enter the school of their choice using scholarships for the 1996-97 school year.

MAY 1997
An Ohio state appeals court holds that including religious schools in the voucher program violates both the state and federal constitutions. Supporters of vouchers appeal; the program is allowed to continue while case is before the Ohio Supreme Court.

AUGUST 1997
Number of students participating in the program rises to 2,938.

AUGUST 1998
Participation rises to 3,774 students.

MAY 27, 1999
Ohio Supreme Court holds that the program violates the state constitution due to a procedural flaw in how the program was enacted. The Court states that the program does not violate the federal Establishment Clause.

JUNE 29, 1999
The Ohio General Assembly reenacts the program in a constitutionally sound way.

SOURCES: School Choice Committee; Center for Education Reform low-income students receive up to $5,553 in tuition relief for private and religious schools. Also likely to fall would be the Florida A+ program, which provides up to $3,472 for children who attend chronically failing public schools. It enrolls only a few dozen students in Pensacola, but has the potential to expand statewide. One might say that for all practical purposes, vouchers would be dead. But the same is not necessarily true of school choice.

> If the Supreme Court decides that vouchers violate the Establishment Clause, it will raise the wall of separation between church and state to a level it has not seen since the Burger Court.

If the Court strikes down the Ohio law, and by implication those in Wisconsin and Florida, choice supporters will probably pin their hopes on funding schemes involving tax relief, like the tax-deduction plan in Minnesota and less ambitious programs in Illinois and Iowa. By providing a more direct benefit to families, tax-relief programs add a level of separation between the state and the school and are generally deemed to be less vulnerable to legal challenge. Activists will also focus their energies on tax-relief programs for third parties that provide scholarships for poor children to attend religious and private schools. Such programs already exist in Arizona, Florida, and Pennsylvania. In fact, more children (60,000) participate in privately funded voucher

JULY 20, 1999
The ACLU, People for the American Way, and the national teacher unions file suit in federal court, alleging that the program violates the separation of church and state.

AUG. 27, 1999
Judge Oliver alters injunction. He allows only previously enrolled scholarship students to return to school. But this leaves 817 new students shut out. Defendants appeal.

NOV. 5, 1999
In a 5-4 decision, the U.S. Supreme Court overrules Judge Oliver's injunction and restores scholarship funding to 817 children.

DEC. 11, 2000
In a 2-1 decision, the federal appeals court rules against the program. The decision contains language suggesting legislative remedies that will satisfy the court.

SEPT. 2001
U.S. Supreme Court agrees to hear the case.

1999 2000 2001 2002

AUG. 24, 1999
Federal district court judge Solomon Oliver Jr. grants a temporary injunction, shutting down the program pending a full hearing.

OCT. 19, 1999
With no response from the U.S. Court of Appeals for the 6th Circuit, defendants appeal to the U.S. Supreme Court.

DEC. 20, 1999
Judge Oliver rules the program unconstitutional and stays his decision, allowing children to remain in school while defendants appeal to the 6th Circuit.

MAY 2001
The state of Ohio appeals the decision to the U.S. Supreme Court. The program continues while the appeal is pending.

FEB. 20, 2002
U.S. Supreme Court hears oral arguments on the constitutionality of the Cleveland Scholarship and Tutoring Program.

The existence of privately funded voucher programs in nearly every state has introduced poor parents to the idea that there is an alternative to failing inner-city schools.

programs than in publicly funded programs. Private initiatives such as the Children's Scholarship Fund and Children First America are likely to grow no matter what the outcome of the Ohio case.

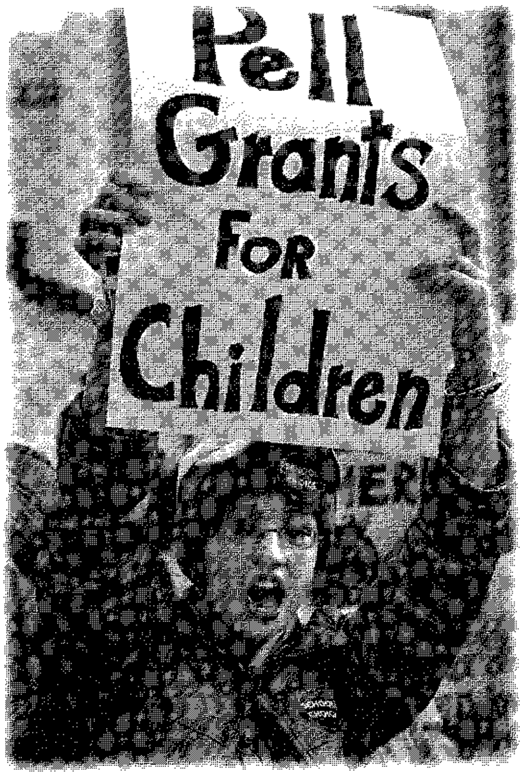
It's Constitutional, but . . .

Suppose the Supreme Court upholds the Cleveland program. Again the impact of the decision depends on the specific questions the majority chooses to address. Will it deem the distinction between direct and indirect aid significant? Will it finally dispose of the legal quibbling over form versus substance regarding the payment procedure? Will it adopt the broader standard of neutrality suggested by Justice Clarence Thomas in *Mitchell v. Helms*? Under this criterion, aid is permissible when it "is offered to a broad range of groups or persons without regard to religion" and "results from the genuinely independent and private choices of individual parents." Or will it decide the case on more narrow grounds? It is quite possible for the case to result in a split majority, as in *Helms*, where four justices accepted the neutrality standard, while Justices Sandra Day O'Connor and Stephen Breyer permitted the aid on the basis of more narrowly defined criteria. Either way there is bound to be another round of litigation.

One possible site for a new legal battle is Maine. Maine has a 130-year-old voucher law that once allowed children living in towns without high schools to attend private or parochial schools with state support. In 1981 the law was changed to exclude religious schools. The shift in policy was upheld by the Maine Supreme Court on First Amendment grounds in 1999 and subsequently confirmed by a federal appellate court, with the Supreme Court refusing to hear an appeal. If the U.S. Supreme Court approves the Cleveland program, choice supporters in Maine could have their case reheard to reinstate the eligibility of religious schools.

It is no accident that most of the recent legal challenges to existing voucher laws began in the state courts. Opponents have based their litigation strategies on "Blaine Amendment" provisions in state constitutions that set a more rigid standard for church-state separation than that enshrined in the First Amendment—at least as the U.S. Supreme Court has interpreted it over the past two decades. Blaine amendments, a remnant of the 19th-century battles over public aid to parochial schools, are named for James G. Blaine, a presidential aspirant and congressman from Maine, who in 1875 tried unsuccessfully to enact a federal constitutional amendment prohibiting such aid. Although Blaine failed to assemble the supermajority of votes needed to pass a federal constitutional amendment, his proposal became a model for state legislators who shared his separationist and anti-Catholic sentiments, then widespread in the nation. By the end of the 19th century, 29 states had written similar amendments into their constitutions. These provisions, as well as others that were added later, could have a major impact on the future viability of voucher programs—but not without opening the door to yet another wave of litigation. Although opponents lost their challenges in the state courts of Ohio and Wisconsin, they prevailed in Vermont, which, like Maine, has a century-old voucher law that disqualified religious schools from participation in 1995. On appeal the Supreme Court refused to review that case also.

The specific exclusion of religious schools from state voucher programs, as in Vermont, raises federal questions beyond the Establishment Clause. Choice supporters claim that such discriminatory exclusion violates the



Federally funded Pell grants may be used to pay tuition at religious colleges. Are they a model for school vouchers?

Characteristics of Voucher Programs for Low-Income Families

A mix of privately and publicly funded voucher programs has spread across the nation, mainly in urban areas.

City or State	Sponsor	Religious Schools Included?	Grades	First School Year	Initial Enrollment	2000-01 Enrollment	Number of Schools 2000-01	Maximum Payment in 2000-01 Dollars	Selection Method
Milwaukee	State of WI	Yes	preK-12	1990-91	341	9,638	103	5,326	Lottery
Indianapolis	ECCT[a]	Yes	K-8	1991-92	746	2,387	82	1,000	First-come[j]
Milwaukee	PAVE[b]	Yes	K-12	1992-93	2,089	819	52	1,000/elem. 1,500/high	First-come
San Antonio	CEO[c]	Yes	1-8	1992-93	930	1,319	62	4,000	First-come
Wash., D.C.	WSF[d]	Yes	K-12[h]	1993-94	30	1,300	116	2,000/elem. 3,000/high	Lottery
Cleveland	State of OH	Yes	K-8	1996-97	1,996	3,900	67	2,500	Lottery
New York City	SCSF[e]	Yes	1-5	1997-98	1,200	1,650	216	1,400	Lottery
Dayton	PACE[f]	Yes	K-12	1998-99	542	680	42	1,785/elem. 2,300/high	Lottery
Florida	State of FL	Yes	K-12	1999-00	146	52	2	3,500	Lottery
Charlotte	CSF[g]	Yes	2-8	1999-00	388	438	52	1,700	Lottery
National	CSF[g]	Yes	K-12[i]	1997-98	1,000	40,000	7,000	1,700	Lottery

[a] Educational Choice Charitable Trust
[b] Partners Advancing Values in Education
[c] Children's Educational Opportunity
[d] Washington Scholarship Fund, Incorporated
[e] School Choice Scholarships Foundation
[f] Parents Advancing Choice in Education
[g] Children's Scholarship Fund. Note: The program was initially organized in Washington, D.C., and was expanded nationally for the 1999-00 academic year.
[h] Students must be in grades K through 8 to begin the Washington Program.
[i] Once awarded a scholarship, all students are guaranteed continued assistance for three additional years. However, the first year's scholarship must be awarded while the student is enrolled in grades K through 8.
[j] Program enrollment in Indianapolis is supplemented with periodic lotteries.

SOURCE: William F. Howell and Paul E. Peterson with Patrick Wolf and David Campbell, *The Education Gap* (Brookings, 2002)

Fourteenth Amendment's equal protection clause and the First Amendment's free exercise clause. Since the state constitutional issue has already been addressed by the Ohio Supreme Court, there is no reason to expect the question to arise in *Zelman*. But if the U.S. Supreme Court determines that vouchers are allowed under the Establishment Clause, it is only a matter of time before the Court will be asked to settle these larger questions. The same Court that set guidelines for permissible aid in *Witters* also left the door open for states to set their own standards for church-state separation. While adopting a more accommodating approach to the First Amendment than its predecessor, the Rehnquist Court has also shown a strong sympathy for state prerogatives on matters of federalism. Inevitably, voucher proponents will insist that the constitutional rights secured by a victory in *Zelman* would prove hollow if the states were permitted to undermine choice on their own. What the Court does to resolve the inevitable clash between these claims and its notion of federalism remains to be seen.

Back to Politics

Even if the Supreme Court were to resolve the federal and state legal questions in favor of vouchers, it would only be setting the stage for the next arena of conflict. Courts only review laws; they do not make them. The most generous judicial interpretation of the voucher question could at most require that states not exclude religious schools from choice programs that are open to other private schools. States would be allowed to continue restricting public funding to government-run public schools, as most do now. Battles over school vouchers have already taken place in more than half the state legislatures, and they will go on.

Schools participating in the Cleveland voucher program represent only a small portion of the range of choices available outside the regular public schools.

No doubt a ruling in their favor from the Supreme Court would reinvigorate voucher proponents. It might motivate President George W. Bush—whose solicitor general gave oral argument in *Zelman* in support of the Cleveland program—to revisit his controversial proposal for federal vouchers. (He has already endorsed tax credits in his current budget proposal.) But the more significant political battles will be fought in the state legislatures, where most education policy is made. Once again the structure of the alliances that form will be filled with political irony. As was so in Cleveland and Milwaukee, the most consistent advocates for school vouchers in America are low-income black and Hispanic parents who live in central cities where the public schools have a history of poor performance. Some black leaders—such as the Reverend Floyd Flake, a former congressman from New York; city councilman (and mayoral candidate) Cory Booker of Newark; and Howard Fuller of the Black Alliance for Educational Options—see choice as a civil-rights issue, a mechanism to provide poor families with the same opportunities enjoyed by the middle class—indeed, as a fulfillment of the promise articulated in *Brown v. Board of Education*: to make education available to all "on equal terms."

The most consistent advocates for school vouchers in America are low-income black and Hispanic parents who live in central cities.

However, the majority of black and Hispanic political leaders oppose vouchers. Their position is supported by mainline organizations like the NAACP, the American Civil Liberties Union, and the National Urban League, all of which have a long history of advocacy on behalf of disadvantaged populations. While reaching out to fellow Democrats, choice proponents in the minority community have sought to build alliances with Republicans and with libertarian organizations like the Institute for Justice, which has represented poor parents in every voucher case that has come before the courts in the past dozen years, including *Zelman*. These are not always easy partnerships. People on the left side of the political spectrum favor targeted choice aimed at the poor, while those on the right prefer universal vouchers made available to all parents. Nonetheless, these alliances have managed to move choice along in places like Wisconsin, Ohio, and Florida.

The Democratic Party has its own tensions to resolve under an ideological tent that tries to accommodate both old-line labor unions, which instinctively oppose school choice, and a younger generation of black and Hispanic activists who demand it. Party leaders have failed to respond adequately to the question of why poor minority parents should be required to send their children to failing public schools when luminaries like Bill Clinton, Al Gore, and Ted Kennedy saw fit to send their own children to private schools.

To a large extent, the choice cat is already out of the political bag. The development of voucher programs in Wisconsin, Ohio, and Florida has fostered a serious national debate over a question that once could be discussed only on the outer margins of politics. More important, the existence of private voucher programs in nearly every state has introduced poor parents to the idea that there is an alternative to failing inner-city schools, and it is winning more converts every day. That being said, polls indicate that the nation as a whole is at best ambivalent about using tax money to send children to religious schools. And voucher proposals are consistently rejected in popular referendums, as in Michigan and California during the 2000 election, where vouchers were defeated by a 2–1 margin. Furthermore, teacher unions that vehemently oppose vouchers are a powerful force within most state legislatures, almost assuring rejection in most places. Yes, the political debate is very much alive, but it remains tilted against choice.

When the Court Speaks

What the Supreme Court says in *Zelman* could have a marked effect in structuring the terms of the political debate—not just in determining who wins the legal argument, but in explaining its broader implications in a way that only the Supreme Court can. In *Zelman*, the Court is being asked to weigh two competing political values: strict church-state separation on the one hand and the right of poor families to choose the education their children receive on the other. If the majority settles on a strict interpretation of the Establishment Clause, it will add an air of legitimacy to an already dominant political coalition that opposes school vouchers and other forms of private-school choice. It will raise the wall of separation between church and state to a level it has not seen since the Burger Court.

The political impact of the court's decision could be even greater if it approves the Cleveland program. At a minimum it would lift the constitutional cover from those political actors who hide behind the First Amendment as a reason to oppose choice. It would lay bare a fundamental struggle over who controls the education of children, the parents or the providers. It also might help to reverse the prevailing political dynamic. Depending on the wording of the opinion, the decision could add a moral dimension to the pleas of poor parents who want educational choices similar to those enjoyed by the middle class. It is difficult to overestimate the power of moral argument in American politics. History has shown it to be an essential ingredient for reversing dominant political patterns in response to demands by weaker parties in pursuit of social justice. The Court provided such a platform in the *Brown* decision in 1954. While the immediate impact of the ruling was to prohibit de jure school segregation, *Brown* breathed life into the political struggle that, against all political odds, brought about a revolution in public policy, affecting every branch of government at the federal, state, and local levels.

> **Fifty years later, the spirit of *Brown v. Board of Education* is vastly more important than its wording. So it may be with *Zelman*.**

Whether the Supreme Court perceives school choice as a fulfillment of the promise articulated in *Brown* remains to be seen. Judicial majorities do tend toward more narrow rulings, except of course when they have something more significant to say beyond the particular legal questions set before them. This may be one of those extraordinary times. If so, such an opinion would be especially compelling coming from the pen of Justice Thomas, the lone black member of the Court, who has written for the majority in a number of important cases involving religion and education. It would remind the nation that there is more than one voice in the black community, each driven by a vision of educational equality, following different paths to get there.

But even a more narrow decision may prove to be pathbreaking. After all, *Brown* itself was a cautious decision, declaring unconstitutional only segregation in schools, not segregation at train stations, parks, or other public facilities. And *Brown* only required the states to implement school desegregation "with all deliberate speed," something less than a clarion call for immediately rectifying the effects of racial injustice. Yet 50 years later, the spirit of *Brown* is vastly more important than its wording. So it may be with *Zelman*.

–Joseph P. Viteritti is a research professor of public policy at New York University and the author of Choosing Equality: School Choice, the Constitution, and Civil Society *(Brookings).*

The Supreme School Board

[illegible] by PAUL E. PETERSON

The waiting line to hear oral argument before the U.S. Supreme Court formed the night before February 20. Anyone joining after 5 A.M. never got in—except those given special seating, including such notables as Senator Edward Kennedy, Health and Human Services secretary Tommy Thompson, and former White House counsel C. Boyden Gray. It was well worth the wait. Persistent questioning, passionate debate, direct self-contradictions, an electric atmosphere—all were there. As the 80-minute conversation came to an end, a pro-voucher resolution seemed to have just barely emerged, the outcome turning as much on educational facts as constitutional questions.

The Court seemed as much a national school board as an interpreter of the Constitution's Establishment Clause. Questions seldom focused on past jurisprudence—probably because earlier decisions have constructed a wall of separation between church and state as serpentine as the one Thomas Jefferson designed for the University of Virginia's campus. Instead, the day's focus was on vouchers, charter schools, and the woeful state of public education in Cleveland. The justices seemed to realize that they were discussing the future of low-income, inner-city children, not just fine points of legal doctrine.

It was Justice David Souter who first posed the central question to Ohio assistant attorney general Judith French: "Isn't it true that something like 99 percent of the students who were receiving these vouchers are in religious schools?" Such restricted choice was very different from the "choice from [among] the great universe of colleges and universities," where federal aid to religious institutions has been generally regarded as constitutional.

To some of the justices, the choices in Cleveland appeared even more restrictive than the 99 percent figure suggests. In their eyes, the high performance of parochial schools relative to the public schools was damning, at least from a constitutional perspective. "The better the parochial school," said Justice Stephen Breyer, "the less the freedom of choice. . . . If it were my children and I saw these comparisons, I'd say, send them to the parochial school . . . That's not my religion, but it's very important my child get the best education, and therefore I would be feeling I had to send them there, if that's what I want."

Yet just as it seemed the Court was about to conclude that parochial schools are so good that school choice in Cleveland is meaningless, charter schools—known as community schools in Ohio—were called on to save the pro-voucher argument. In the words of Justice Antonin Scalia, "I assume Justice Breyer could send his child to one of the community schools, which [are] entirely nonsectarian. . . . [These] schools get more money than the sectarian schools." Scalia also noted that the more established and better funded voucher program in Milwaukee has been attracting secular schools in steadily increasing numbers.

When it came time for the anti-voucher forces to make their defense, Robert Chanin of the National Education Association stepped forward. (Curiously, when veteran reporter Linda Greenhouse's story appeared in the *New York Times* the next day, it failed to state Chanin's NEA connections, identifying him only as the attorney for the Cleveland residents who had challenged the program.)

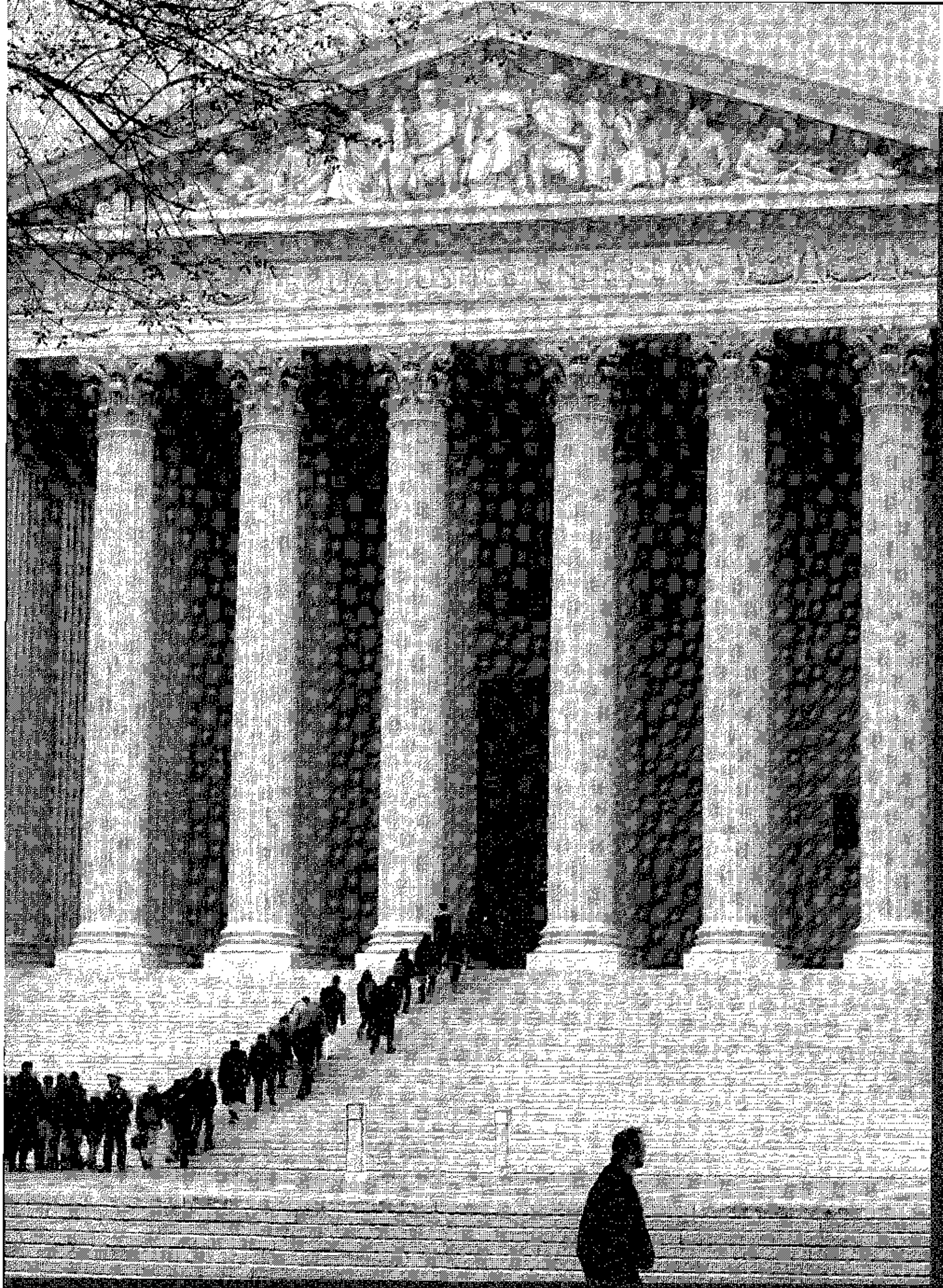
Chanin's most difficult task was to show that the community schools in Cleveland were irrelevant because, in Chanin's view, the justices were legally required to look not at the entire situation in Cleveland but only at the specific statute creating the voucher program. When the attorney restricted his legal vision in this way, he was able to argue, "It is a mathematical certainty that almost all of the [voucher] students end up going to religious schools."

No sooner were these words enunciated than the most dramatic moment of the morning arrived—a clear, decisive intervention by Justice Sandra Day O'Connor. It was not just what she said—though this was powerful enough—but the fact that O'Connor is expected to cast the decisive vote in this case, as in so many others. For months, even years, it has been evident that the outcome could easily turn on her vote—and the scope of the opinion might well depend on her views. Justices Souter, Ruth Bader Ginsburg, and John Paul Stevens were unlikely to find the law constitutional, and Breyer may have tipped his hand when he observed, rather whimsically, that parochial schools may become increasingly unconstitutional the more they outshine their public-school counterparts. Meanwhile, pro-voucher groups are taking heart from the comments—as well as previous opinions—of Justices Scalia, Clarence Thomas, Anthony Kennedy, and Chief Justice William Rehnquist.

The justices themselves were keenly aware of O'Connor's decisive position. Both sides made subtle appeals to her. Said Souter: "What's bothering me . . . and, I suspect, O'Connor, too," is that the law must be not only neutral on its face but also in its effect, and "at the end of the day, the effect is a massive amount of money [going] into religious schools. . . . That is the sticking point here."

O'Connor had indicated that the voucher program might resemble a New York State tuition-reimbursement program struck down in the 1973 *Nyquist* case. Though this observation must have given hope to the NEA attorney, when he began insisting on mathematical certainties, he encountered tough, if patient, resistance: "Well, wait just a minute," said O'Connor. "Do we not have to look at all of the choices open to the students, the community schools, the magnet schools, et cetera?" Chanin did his best: One must legally ignore all the other schools in Cleveland because "this court has always been program-specific in its financial-aid cases."

"But I'm not sure that's proper," replied O'Connor. "That's what I'm asking you. Why should we not look at all of the options open to the parents?" Doing so, Chanin argued, "mixed together programs



that are quite qualitatively different in both function and purpose." But, persisted O'Connor, "is it not true that parents can choose to have their children educated in a community school and, if they do, that school gets more money from the State than if they had chosen the religious school? If anything, it's skewed against the religious schools." When Chanin iterated, "We now have this year 99.4 percent of the students going to religious schools," Justice Kennedy, making his own appeal to O'Connor, dropped the acid observation: "So far, you're doing a very good job of not answering Justice O'Connor's question." Laughter temporarily broke the tension in the air.

But within minutes, Chanin was in trouble again: "Supposing there are 10 schools out there, 10 private schools, nine of which are nonreligious and one of which is religious," imagined Chief Justice Rehnquist. "Is that . . . consistent with the Establishment Clause?" Replied the NEA attorney: *"Oh, that's clearly* unconstitutional, Your Honor." Pressing hard, Rehnquist observed: "The interesting thing . . . your view is, if any one [religious] school gets the money, it's unconstitutional?" "No, no, your honor," replied the hapless attorney. "Oh, I thought you said yes," said the Chief Justice. "No, I'm sorry if I—I did not," Chanin replied. "Or, I may have, but I didn't mean to."

Though the drama then began to subside, the Supreme School Board still had educational points to make. Late in the morning, Justice Kennedy asked: "Is it unconstitutional for [the State of Ohio] to . . . have a structure in which different school systems, different curriculums, curriculums that do not inflict terminal boredom on students, can begin to flourish? And . . . you say they cannot do it?"

Chanin could only reply: "There is no evidence that competition improves the lot for the 96 percent of the students who remain in the troubled Cleveland public school system with less resources and even worse problems." Even this did not remain unchallenged. "The studies that I'm familiar with say that the inner-city parochial schools, which spend much less per child on education, do a better job than the public schools that spend much more," said Scalia, adding "so I just don't think it follows that . . . more money [will] solve the difficulty that the people of Cleveland found with their public schools."

When told that the only evidence about parochial schools was anecdotal, Scalia added, "Oh, I don't think it's anecdotal at all. I mean, there are extensive studies that show the parochial schools do a better job."

Still, questions posed in oral argument do not necessarily translate into opinions on judgment day. But if the tenor of the discussion on February 20 is an indicator of what is to come, then the future of school choice will take a new twist. If Cleveland's charter schools are the key to making vouchers constitutional, must future voucher schemes also include a charter component? And must the vouchers be large enough that, as in Milwaukee, they invite increasing participation by secular schools? If vouchers are found constitutional only if charters are available and secular private schools open themselves to voucher recipients, the result could profoundly affect the future of school choice in ways neither side anticipated. The Court may turn out to be the Supreme School Board in deed as well as in the words it voiced in the oral argument.

–Paul E. Peterson is the editor-in-chief of Education Next *and co-author of* The Education Gap *(Brookings, 2002).*

feature

by SIOBHAN GORMAN

Bipartisan Schoolmates

President Bush deserves credit for forging a consensus on federal education policy. But it will all be for naught if the law is not enforced

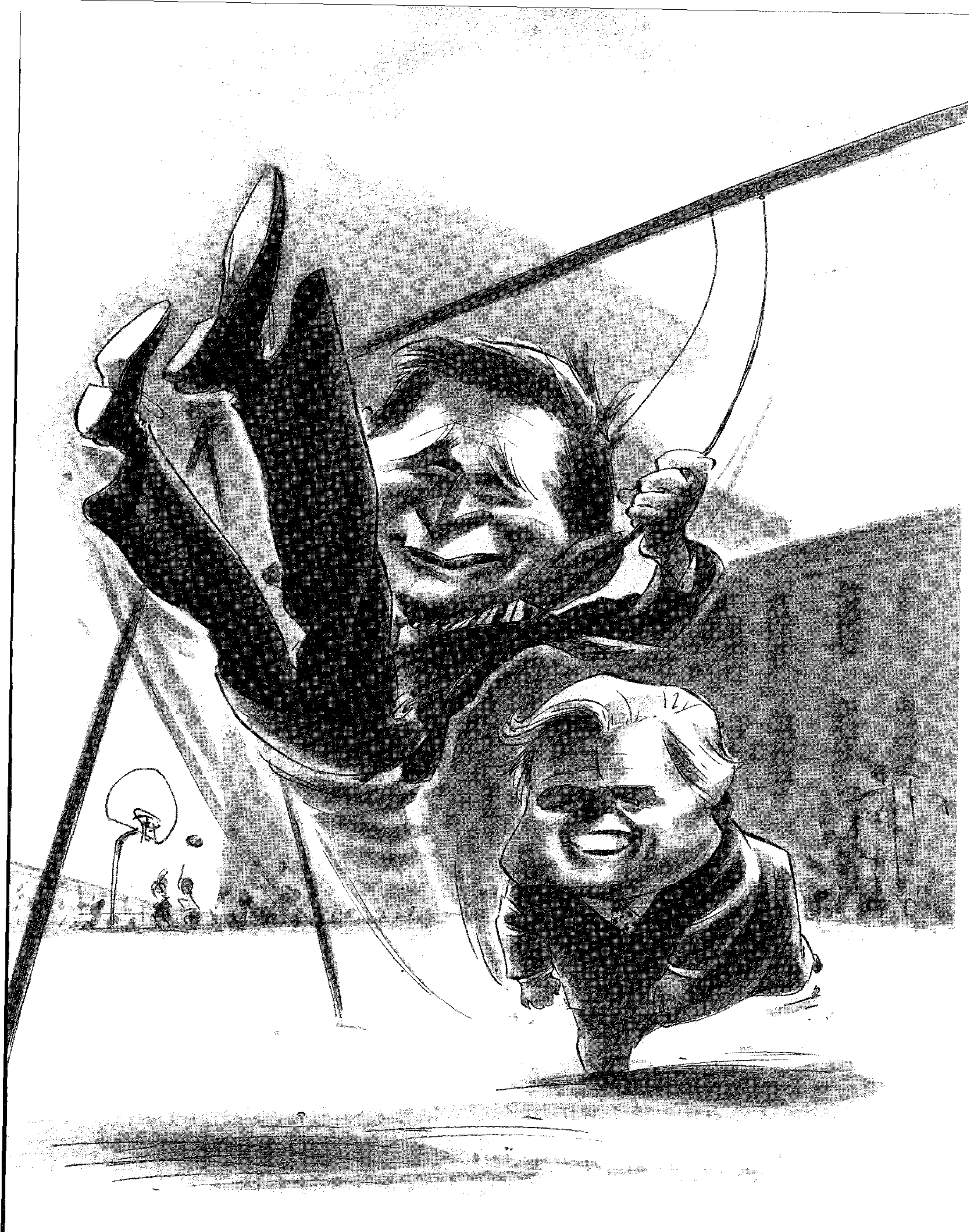
AMID A THRONG OF CHEERING TEENAGERS IN THE Hamilton High School gymnasium, President Bush declared victory. "Today begins a new era, a new time in public education in our country," he intoned confidently into the microphone. "As of this hour, America's schools will be on a new path of reform, and a new path of results." With that, he signed into law the bill that reflected much of his education agenda. This Ohio school was the first stop of the two-day tour the president and his merry band of liberal and conservative lawmakers took in January to celebrate the president's first true bipartisan triumph.

But the party's a bit premature. To be sure, the president gets credit for forging a consensus after years of ideological and political gridlock. And the law does make some significant changes in federal education policy. Nevertheless, the reauthorized Elementary and Secondary Education Act (ESEA) hardly represents a new era. Instead, it builds on at least a decade's worth of federal reform efforts. Where the Bush administration could make its mark is in its *enforcement* of the law, an area in which few, if any, previous administrations have found the political will to play rough with the states.

The law calls for annual tests in reading and math for children in grades 3 through 8, plus a science test in three different grade levels by the 2007–08 school year. (States were already required to test students once in high school.) States must also establish a definition of a failing school that meets federal guidelines. Schools labeled failing for two or more years face increasingly stringent penalties, which states must impose. The law also leaves a host of issues unresolved, giving the states and the federal Department of Education plenty of wiggle room. Thus the quality of implementation may vary widely.

The federal government sports a sobering track record when it comes to enforcing its education reforms. President Clinton's 1994 ESEA reauthorization was hailed by proponents as "the most extensive revision of the legislation since [its enactment in] 1965." It required the states to develop standards and assessments linked to the standards. But lawmakers like Representative George Miller, D-Cal.,

ILLUSTRATION BY DAN VASCONCELLOS



say the changes envisioned by the 1994 reauthorization didn't live up to their billing because "its implementation was fudged by the administration."

Lawmakers in both parties hope that this time a combination of tough federal sanctions, more public reporting of student performance, and an aggressive White House will be enough to prompt change in the schoolhouse. "This bill delivers the goals and the tools to achieve them," says Miller, one of the four top congressional negotiators on the bill. Still, he says, "The bill is not a silver bullet."

A Long and Winding Road

President Clinton kicked off the ESEA debate in May 1999, when he submitted his proposal to reauthorize the law, which was due to expire that year. But Congress continued arguing about the proposal into 2000, and it was shelved in the din of the presidential election. With the two candidates running even in the polls, neither party wanted to risk passing a bill when there was a good chance that they could send a bill that was more to their guy's liking the next year. Furthermore, education would be a less potent issue on the campaign trail if Congress passed a major education reform bill.

Once elected, President Bush sent the early message that he wanted an education bill, and he wanted it to be a bipartisan one. Before his inauguration, Bush held a bipartisan education meeting in Austin, and among the hand-picked Democrats attending was Sen. Evan Bayh, D-Ind. Bayh and Sen. Joseph I. Lieberman, D-Conn., had put forth an education reform bill in early 2000 called the Three R's, aimed at bridging the partisan divide in education.

Borrowing liberally from Lieberman and Bayh's reform package, Bush said that the 54 federal elementary and secondary education programs should be consolidated into five categories reflecting federal priorities: 1) educating disadvantaged students; 2) teacher quality; 3) English fluency; 4) school choice; and 5) school safety. Bush also proposed that states begin testing children in grades 3 through 8. He wanted to allow children in schools that failed to close the achievement gap for three straight years to use federal money to attend private schools. On his second full day in office, Bush unveiled an education "blueprint" that was essentially the same as his campaign proposal. The White House immediately began negotiating with the 10 centrist Democrats cosponsoring the Lieberman-Bayh bill, which they reintroduced the day Bush unveiled his proposal, in an effort to cut a deal quickly.

Once elected, President Bush

However, when it became clear by the spring that Bush's tax cut would pass with the help of a few centrist Democrats, the Lieberman-Bayh group became leery of ditching their party a second time. The White House had also begun negotiating with

No Child Left
No Child Left

Lawmakers in both parties hope that this time a combination of tough federal sanctions, more public reporting of student performance, and an aggressive White House will be enough to prompt change in the schoolhouse.

PHOTOGRAPH BY AP/WORLD WIDE PHOTOS

Sen. Edward M. Kennedy, D-Mass., and the centrist Democrats feared that the White House would play the centrist and liberal Democrats against one another. For instance, the White House got Kennedy to agree to a limited voucher to pay for tutoring services. It then took that agreement to the New Democrats, who had been holding out on a voucher compromise, to try to get the New Democrats to incorporate the tutoring proposal into the deal they were negotiating separately with the White House. To avoid this, Lieberman decided that future negotiations would have to include Kennedy.

sent the early message that he wanted a *bipartisan education bill.*

A new negotiating group formed, this time including the White House, Republicans, New Democrats, and Kennedy sympathizers. After a month of negotiations, the group reached agreement on the two most controversial issues: vouchers and block grants. They went with Kennedy's voucher compromise, which allowed students in failing schools to use federal money for private tutoring. They also agreed to a scaled-back block-grant proposal that would have allowed 7 states and 25 school districts to sign a performance contract with the federal government that would free them of most federal education regulations in exchange for a promise to improve student performance. They thought they had a bill ready to send to the Senate floor.

But at a late-night meeting in mid-April, one congressional aide announced that 80 to 90 percent of the schools in states like Texas and North Carolina, both of which had seen rising achievement scores through the 1990s, would be deemed "failing" under the bill's definition. The bill required states to set performance goals for every demographic group of students. The problem was that if a school didn't meet that goal for any one group in any grade level in any one year, it would be labeled failing.

Two weeks later, the senators settled on a complicated formula that required states to calculate an overall performance grade for a school based on several factors, including improving test scores for poor and minority children. But no longer could a school receive a failing grade solely because its poor and minority students didn't see their test scores rise. By the time this definition was devised, it was early May. The Senate began debate, but it dragged on for six weeks as the Senate juggled campaign finance reform and turned over to Democratic control.

On the House side, the bill moved more quickly. A series of bipartisan negotiations between Reps. Miller and John Boehner, R-Ohio, chairman of the House Committee on Education and the Workforce, produced a bill that was similar to the Senate's, but without the block-grant provisions. The Republican leadership and a handful of rank-and-file conservatives protested. But President Bush weighed in on Boehner and Miller's side, and the House passed the bill on May 23. The Senate passed its version a few weeks later.

Still, little was settled on the accountability front. The Senate's rejiggered formula was widely seen as too complicated for parents to understand, and the House's formula would run into the same reality-check problems that the Senate's had before it was revised. The conference committee charged with resolving differences between the House and Senate versions of the bill began meeting in late July, but the bill languished in committee as the members squabbled over details small and large.

In the meantime, fears over the shrinking surplus were beginning to dominate debate on Capitol Hill. Democrats charged that Bush's tax cut made it impossible to fund education adequately. Interest groups concerned with some of the testing and accountability requirements began circling.

"There weren't a lot of bipartisan feelings," remembers Bush education adviser Sandy Kress. "[Bush] was concerned. . . . The momentum clearly had slowed. He was aware of the mood and the difficulties when we came back from the August recess. We were supposed to be farther along."

Then September 11 intervened. When he first heard of the terrorist attacks on the World Trade Center and the Pentagon, Bush was visiting a Florida elementary school to publicly prod Congress to send him a final bill. Perversely enough, the attacks probably helped get the process back on track by ending the political bickering, congressional aides said. The negotiations were kicked up to the conference's top four members: Reps. Miller and Boehner and Sens. Kennedy and Judd Gregg, R-N.H. The thorniest issue—the definition of a failing school—was resolved by the Big Four by late September.

Under the new definition, states must design a plan to raise the children in several demographic groups—black, white, Hispanic, poor, and disabled—to a "proficient" level of achievement on state tests within 12 years. States determine what a proficient score will be and choose an initial goal for the percentage of students in each group that will attain a proficient grade that year. They must raise that bar over time, so that each state reaches 100 percent proficiency 12 years later. If a school does not meet the performance goal for one demographic group, but reduces the number of children who are not proficient by 10 percent, the school will avoid federal penalties. Schools that don't meet either goal will be labeled failing.

While congressional negotiators had agreed to this compromise in September, they didn't announce it until the final conference bill was set in order to avoid opening it up to too much criticism from interest groups and others (as had happened in August) and jeopardizing other negotiations in

At a late-night meeting, one congressional aide announced that 80 to 90 percent of the schools in Texas and North Carolina would be deemed *"failing" under the Senate's education bill.*

progress. "We kept it quiet for a long time," said one lawmaker involved in the negotiations process. "It was the most sensitive part of the bill."

Following in Dad's Footsteps

The new testing requirement exemplifies Bush's tough-love approach to education reform. "I understand taking tests aren't [sic] fun," the self-proclaimed C student at Yale University told the crowd at Hamilton High. "Too bad." One of the least debated provisions, the requirement for annual testing, is probably the biggest change in the 2002 law. Though the numbers vary depending on how you read the new law, 15 states currently meet the annual testing requirement in math, 17 in reading, and 24 have established a science test, according to the Education Commission of the States.

This is not, however, a sea change in federal policy. Annual testing builds on the efforts initiated by the Clinton administration in 1994, which in turn built on ideas put forth by President George H. W. Bush when he convened the state governors at the first National Education Summit in 1989. "It's a continuation of an era," said John F. "Jack" Jennings, director of the Center on Education Policy and a longtime Democratic aide on the House education committee.

In 1994, Congress began requiring states to establish academic standards in each grade and to create tests to assess whether children have learned the material. Those tests had to be administered to all poor children at least once in grades 3 through 5, 6 through 9, and 10 through 12. Clinton's 1999 proposal would have expanded that testing to all students. The new law mandates that all students take tests that measure their progress against state standards every year in grades 3 through 8.

The bill Bush signed also requires that the National Assessment of Educational Progress (NAEP) be given in every state every other year in math and reading, but the results cannot influence whether a school is designated as failing. Initially, Bush had proposed that NAEP be administered every year as a national barometer for the state tests. Currently, NAEP is administered in those subjects every four years in about 40 states, a number that fluctuates from test to test depending on which states decline to participate.

The 1994 legislation only suggested penalties for schools that failed to improve student test scores, but Clinton's 1999 proposal, which never made it into law, would have required that states intervene in perpetually failing schools. "The Bush administration took the Clinton administration's ideas and ran with them," said one bemused Democratic congressional aide who worked at the Department of Education under Clinton.

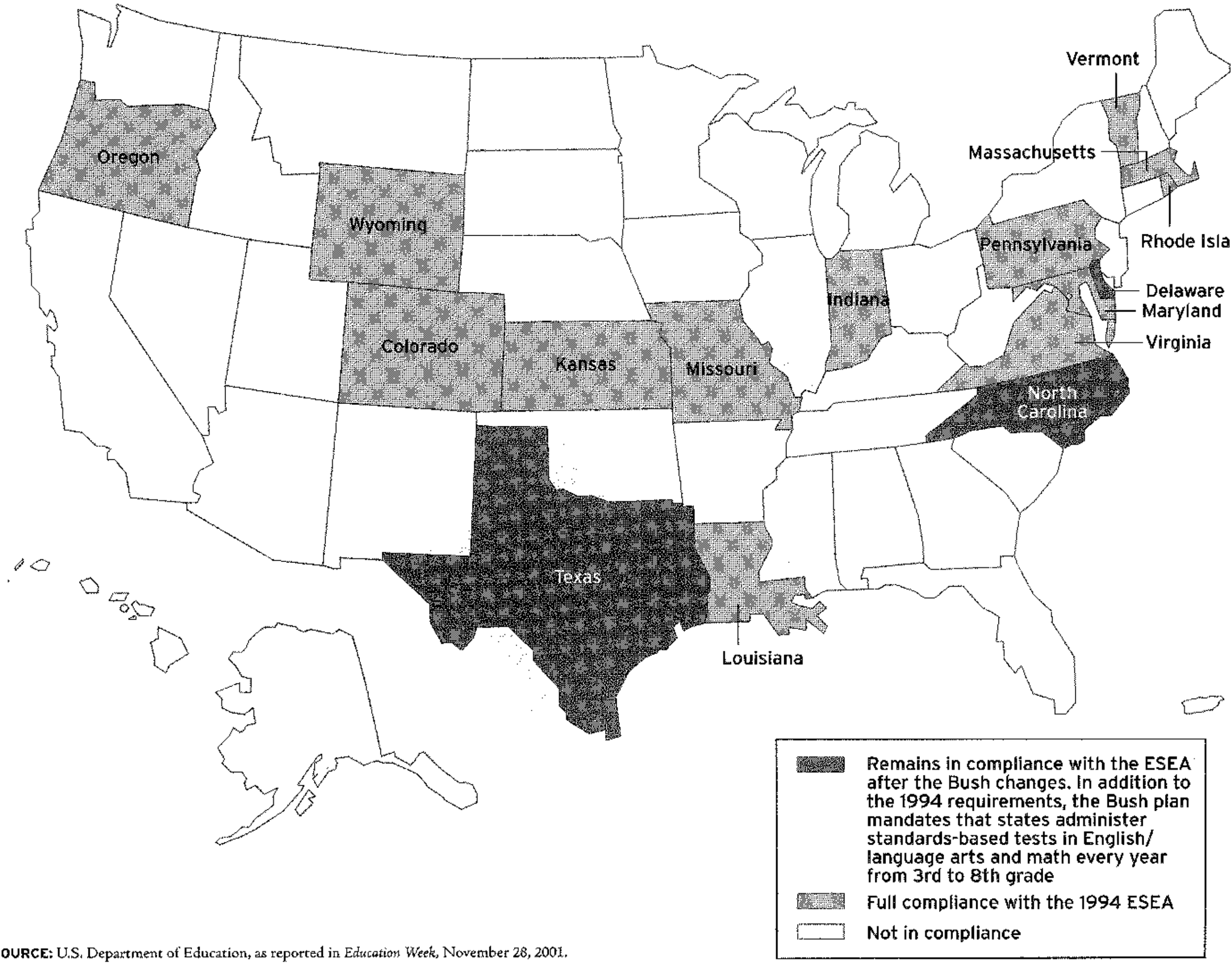
The new law sets out a timeline of increasingly severe sanctions to nudge recalcitrant schools along. If a school fails to meet annual state test-score goals for two years, students can transfer to another public school in the district. The failing school also receives extra money to revamp academic programs. After three years of failure, the district must use 5 percent of the money it receives for poor children under the federal Title I program's "Basic Grants" section to pay for outside tutoring services. After four failing years, a school must make significant structural changes, such as revamping the curriculum or firing staff, and may be eligible for a state takeover the following year. At no point does a state or school lose money if it continues producing poor results, although districts eventually have to divert federal money to pay for tutoring services.

One of the major worries is that, with 50 different statewide tests, the federal government will lack the ability to *compare test scores across state lines.*

Building on the work of one's predecessors can produce success stories. That's just what Bush did as governor of Texas. His reforms expanded those of former governor Ann Richards, whose reforms built on those in places like Dallas and previous statewide efforts led by, of all people, Ross Perot. Dallas, it happens, is where then–school board president Sandy Kress initiated accountability—using standardized tests to reward and sanction schools. "This is our theme," Kress said. "We may not know a lot of music in Texas, but we can sing the song we know well."

Lax Enforcement (Figure 1)

Eight years later, just 16 states are in full compliance with the requirements of the 1994 reauthorization of the Elementary and Secondary Education Act.


Oregon
Wyoming
Colorado
Kansas
Missouri
Indiana
Texas
Louisiana
Vermont
Massachusetts
Rhode Island
Pennsylvania
Delaware
Maryland
Virginia
North Carolina
Remains in compliance with the ESEA after the Bush changes. In addition to the 1994 requirements, the Bush plan mandates that states administer standards-based tests in English/language arts and math every year from 3rd to 8th grade
Full compliance with the 1994 ESEA
Not in compliance

SOURCE: U.S. Department of Education, as reported in *Education Week*, November 28, 2001.

A New Enforcement Era?

The Bush administration tried to set a new tone the night after the president signed the education bill, when Secretary of Education Rod Paige invited 30 state education chiefs to Mount Vernon to discuss the new law and his expectations for the states. After dining on pork and pumpkin mousse, Paige made his pitch: The old days of waivers and delays and closing our eyes to enforcement problems are over.

Analysts say the tough-love message is a good one, but that backing it up will be difficult. "Everybody jumps on this wagon and says we're really going to be tough," said Marshall S. Smith, the acting deputy secretary of education in the Clinton administration. "That's macho talk. That's not going to work in the real world of kids and schools."

So far, only 16 states have complied with the requirements of the 1994 law. The Bush administration's first task will be to get the other 34 states into compliance. Then, they'll have to review the new statewide plans for getting every child in every demographic group to a "proficient" level of achievement within 12 years. After the states have set their yearly performance goals, the Department of Education will have to monitor the performance of the nation's 92,000 public schools to see whether each demographic group in each grade being tested is meeting the state performance goals.

The law also leaves open a number of questions, mostly surrounding the quality of the reforms states undertake. One of the most significant ambiguities is whether states can develop an annual testing system by piecing together state and district tests given in different grade levels. Some critics worry that states will create systems where a student's results on a district test

The new testing requirement exemplifies President Bush's tough-love approach to education reform. This is not, however, a sea change in federal policy. Annual testing builds on the efforts initiated by the Clinton administration in 1994.

PHOTOGRAPH BY AP/WORLD WIDE PHOTOS

one year cannot be compared with results on a state test the next year. "There's going to be a lot of discussion about that," Kress said. "There's no easy answer right now."

Others worry that with 50 different statewide tests, the federal government will lack a meaningful tool with which to compare test scores across state lines. "What happened was you had a whole bunch of folks who wanted standards but not standardization," said Arnold Fege, president of Public Advocacy for Kids, a nonprofit public education advocacy group. "As a result, we've got 49 different states with 49 different standards and 49 different testing systems. The question is what does it all mean? You can't compare the states." (The number is only

49 because Iowa declined to create statewide academic standards per the 1994 law's requirements.)

For many of the testing quality issues, the federal government is taking a hands-off approach. "Our job will not be to pass judgment on the quality of state standards or the quality of the assessments based on those standards," said Undersecretary Eugene Hickok, the Department of Education's point person on enforcement.

The department's approach to enforcement is largely preventative. Hickok is setting up an implementation team in Washington that will be bolstered by staff members in the department's regional offices. The team will assemble an analysis of where states are in the reform process, and it will be updated as states make additional changes. Meanwhile, top department officials will reach out to state officials to encourage them to comply. "Our goal is to respond as quickly as possible, and not just to say yes or no, but to help them achieve it," Hickok said.

Some states and schools are likely to test how serious the administration is about enforcement. "There are going to be a lot of schools that take a wait and see attitude with compliance," said David Shreve, senior committee director for the National Conference of State Legislatures. "I think there will be a great number of states and systems that will say, 'Let's do as much as we can without breaking our backs,' knowing full well that the bill will be up for a new reauthorization before any of these really bad consequences kick in."

States are also now facing budget deficits—at least in the near term. Some liberal members of Congress, like Sens. Paul Wellstone, D-Minn., and James M. Jeffords, I-Vt., even voted against the bill because they view the funding level as insufficient to help the reforms really take hold. Jeffords called the bill "counterproductive, if not destructive." Even Miller, who advocated accountability before it was cool, worries about states' being able to pay to fix the problems the test scores highlight. "It comes in a down economy, and we worry the states are going to be forced to make cuts that will challenge the success of this bill," he said. Just a month after Bush signed the education bill, with Miller and Kennedy at his side, the two Democrats held a press conference to denounce Bush's 2003 education budget, which they say falls way short of what's needed to fulfill the mandates of the new education law.

To a large degree, says Hickok, the key to enforcement will be parents. He hopes that parents of children in failing schools, armed with information about how their schools and children are doing, will force schools to offer them the options that are laid out in the federal law. Having been there and done that in Washington, Marshall Smith agrees. "The only way you're really going to get accountability—accountability at the point where politicians really feel the pressure—is bad press," he said. But, he says, the statistics the federal government is requiring schools to report offer parents no useful information about what their children do and don't know.

The first test of the Bush administration's enforcement chops will come in handling the schools that are already on the Department of Education's list of failing schools. Students in schools that have been on the list for three or more years, estimated to be at least 3,000, should now be entitled to tutoring services, and students in at least 6,500 should earn the option to transfer to another public school.

Once the statewide accountability systems are in place, enforcement will get increasingly difficult as states are required to raise the goals for student achievement no less than every three years. As more and more schools face sanctions, how will the administration handle the inevitable logistical and political quicksand? What happens if all the high-performing schools in a district are full and can't take the children who were granted public school choice because their own school failed two years in a row? "We've had a lot of discussion about it," Hickok said. "The law does not impose the Department of Education into the state to make sure they're doing public school choice the way we want to or supplemental services the way we want to." In fact, Hickok isn't even sure what the federal government can do if students in a failing school can't transfer to another public school in the district because there's no room at the better schools.

Sustaining the momentum on enforcement **may be the Bush administration's biggest challenge.**

Sustaining the momentum on enforcement may be the administration's biggest challenge. "Paige is saying he won't give any waivers. Now we shall see. I hope that he can keep the pressure on," says Jack Jennings of the Center on Education Policy. "Elections come up. Is Bush going to deny [New York] Gov. [George] Pataki federal aid when he's running for reelection, if the state has not met its testing requirements?" (Money can be withheld from a state for not complying with the requirements of the law, but not for poor results.)

Even with a troubled economy and continuing worries about terrorism, education remains a high-level concern among voters. Bush's education record will no doubt play an important role in the next election. That means the person who will ultimately be held most accountable in the eyes of the public for reforming the schools is President Bush.

–Siobhan Gorman covers education for National Journal.

Swing STATE

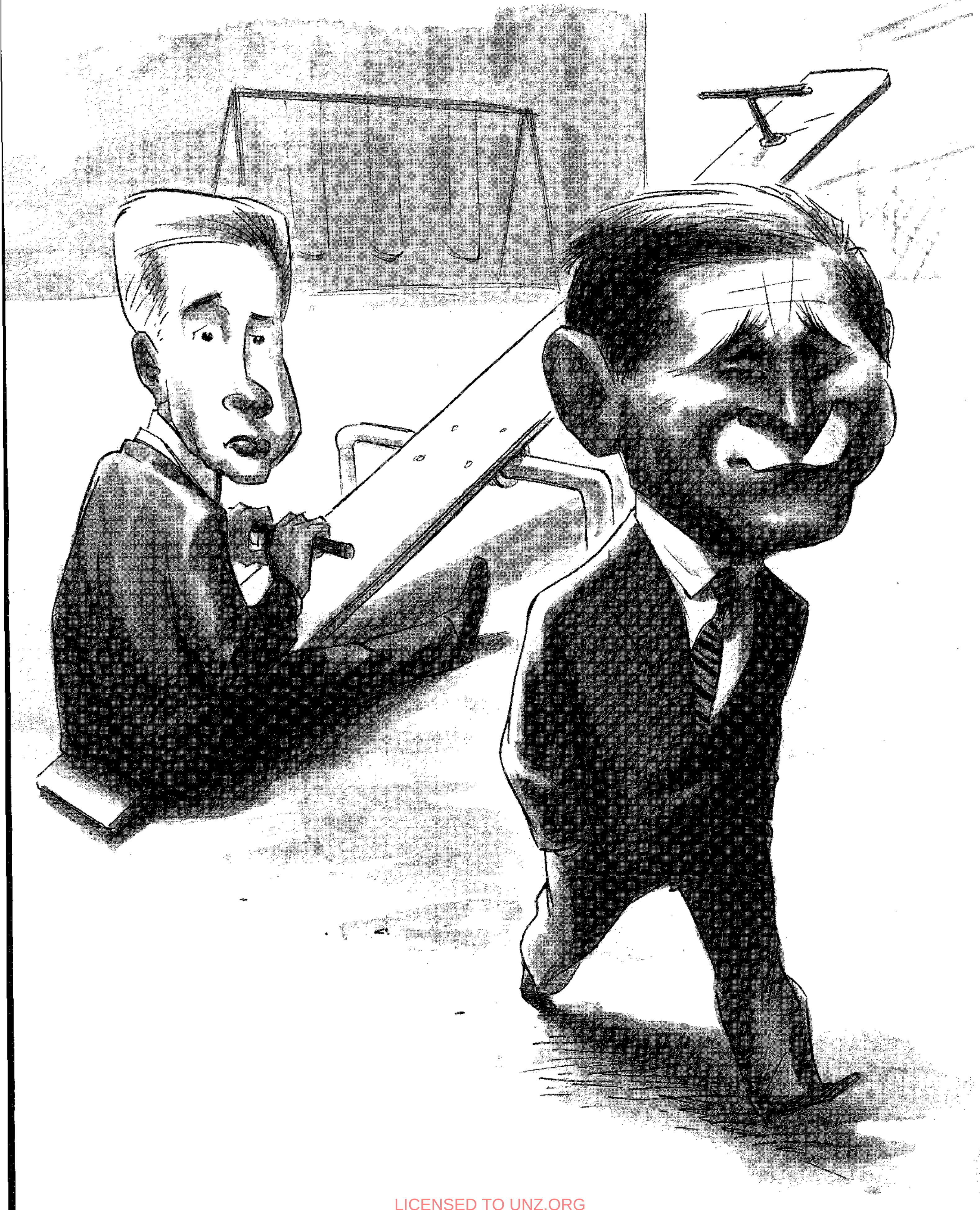
The downs and ups of accountability in California. Will the state's political leadership stay the course this time around?

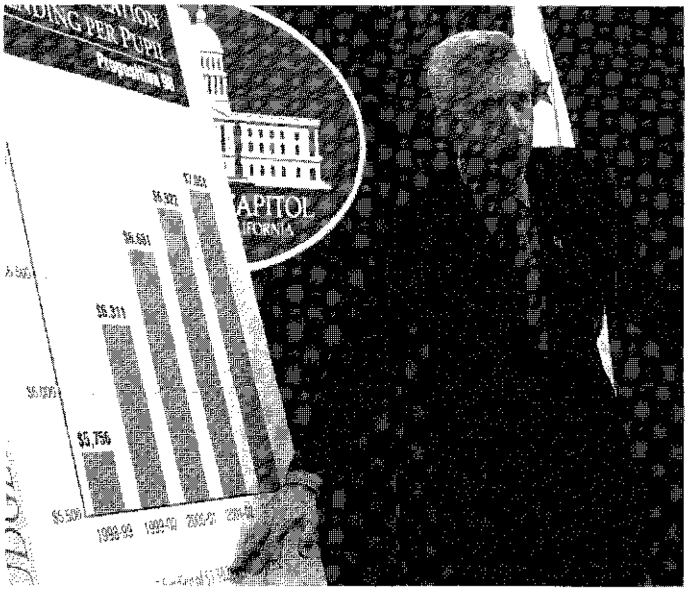
by MICHAEL W. KIRST

CALIFORNIA, NORMALLY A BELLWETHER STATE, ITS trends and culture spreading as fast as buzz on the latest Steven Soderbergh flick, was a late arrival to the modern school-accountability movement. Even the 1994 federal Title I reforms, which required states to develop the three major prongs of an effective accountability system (academic standards, tests linked to the standards, and a mixture of assistance and sanctions for low-performing schools) did little to stimulate California into action. In fact, it wasn't until the results of the 1994 National Assessment of Educational Progress (NAEP) in reading were released that the state got serious about accountability. California's absolutely disastrous performance—it tied Louisiana for last place among the 37 participating states—was a source of deep embarrassment to a state that had long prided itself on its K–12 public schools and the University of California's eight undergraduate campuses, arguably the finest system of public universities in the world.

Analysts have cited a legion of reasons for the state's slide in achievement: the steady leaching of resources from the schools that was the inevitable result of the infamous 1970s property-tax revolt led by Howard Jarvis; a long period of economic woes caused by layoffs in the defense industry; curriculum experiments with "whole language" reading instruction and "new math" that were at best a distraction and at worst quite damaging; a school finance lawsuit that led to a dramatic increase in the state's authority over school budgets and operations; and a massive influx of new students and non-English-speaking immigrants that almost surely depressed test scores. Whatever the reasons,

ILLUSTRATION BY DAN VASCONCELLOS





California's successions in leadership and shifts in the political winds have led to the dismantling of earlier reforms and the layering on of new ones.

PHOTOGRAPH BY AP/WORLD WIDE PHOTOS

the result was a sharp drop in the public's and policymakers' confidence in the abilities and, indeed, the motivation of local educators. A political coalition and consensus developed around the ideas that accurate information on student and school performance is needed in order to hold educators accountable and that educators can't be trusted to work hard without the existence of positive and negative state incentives.

Since then, California has developed the basic foundations of a coherent accountability system. Nevertheless, despite the progress, many aspects have yet to fall into place at the school level, and there is always the danger that the underlying political coalition will collapse and take the accountability system down with it. Education reform generally, and in California especially, seems to follow a pattern of taking one step up and two steps back, as successions in leadership and shifts in the political winds lead to the dismantling of earlier reforms and the layering on of new ones. This is certainly true of the accountability movement in California, which has a tortured history of reform and retrenchment that for years left the state with no real measure of how its students were performing. So the question is, Will California stay the accountability course this time? Will the system survive the almost inevitably low passage rates that will occur when the state shifts to a new high-school exit test this year? Will the state not only punish low-performing schools, but also give them the resources and technical assistance necessary to build solid academic programs? And will the new provisions of the federal Elementary and Secondary Education Act (ESEA) hamper or support California's efforts to improve schools?

Experiments Gone Awry

California may be a latecomer to the modern accountability movement, but the state has been no stranger to the idea that standardized tests ought to be used to gauge school performance. California state officials began to lose confidence in local school authorities and teachers as early as the 1960s. This concern culminated in the California Assessment Program (CAP), which was first administered in 1972. The program required each student to take a sample of questions from an overall test, a method called "matrix sampling." Matrix sampling did not produce scores for individual students, but it did permit an in-depth assessment of a school's performance in each subject. Each school was given a grade based on a comparison between its actual scores and its predicted scores, which in turn were based on its students' socioeconomic backgrounds. Many newspapers published the schools' grades.

In the 1980s CAP became the victim of a political dispute between Republican governor George Deukmeijian and the elected state superintendent of education, Bill Honig, a Democrat. Honig signaled that he was considering running for governor against Deukmeijian, which led to the governor's vetoing of CAP. Local educators did not mourn the death of CAP, since many local schools had received negative publicity as a result of their low test scores. They also contended that CAP did not assess what teachers were teaching. The effectiveness of CAP relied on the willingness of parents, educators, and school boards to use information as a tool for school improvement. Policymakers especially hoped that angry parents would exert pressure on low-performing schools and school boards, but their hopes were never realized.

It wasn't until a decade later that Californians again had detailed information on the state's academic performance. In the early 1990s, Honig created the California Learning Assessment System (CLAS). While CAP had focused exclusively on multiple-choice tests, CLAS asked students to read a poem or passage and respond to questions like: "Pick a part that is especially interesting and explain your reasons," or "What are your feelings about this poem?" Educators, teachers especially, liked the test's open-ended and creative nature, but rumors quickly spread among some parents that the tests contained "objectionable content" that threatened moral values and students' privacy. Other critics questioned whether open-ended items could be objectively scored and thus serve as a reliable measure of school performance. What bothered Republican governor Pete Wilson, however, was that CLAS didn't provide scores for individual students. Wilson believed that individual test scores enhanced parental oversight and responsibility for their children. In the end, despite the support of many educators, CLAS could not overcome its technical and political problems—state superintendent Honig, the major supporter of CLAS, was convicted of a felony and removed from office—and it succumbed to Governor Wilson's veto in 1995.

Californians were again plunged into darkness, as the state returned to allowing each school district to choose its own test. So it came as quite a surprise when the state's vaunted school system finished lower than Mississippi's on the 1994 NAEP. State policymakers, led by Governor Wilson, reacted with alarm and passed a new state standards and assessment program in 1996, called the Standardized Testing and Reporting system, or STAR. New procedures that gave the governor and other public representatives more influence over test questions were designed to overcome the concerns that had engulfed the CLAS program. A commission was established to develop "academically rigorous" standards in all major subject areas, at every grade level. The majority of the commission's members were appointed by the governor. A six-person statewide review panel was to review all test items to ensure that they were free of questions about:

- a student's or parent's personal beliefs, sex life, family life, morality, or religion; or
- personal characteristics such as honesty, integrity, sociability, or self-esteem.

One result is that the new California Standards Tests (known as the CST), which were first administered in the spring of 2001 but won't enter the accountability system until 2002, rely only on closed-end multiple-choice questions, unlike the CLAS test.

The 1996 law envisioned a logical sequence of events. First the state would create and approve curriculum standards in each grade, and then a set of assessments would be developed that were linked to the state's curriculum. However, California's policymakers, concerned as they were about the lack of performance information and accountability in the system, decided that the state couldn't wait that long. So, starting in 1997, all students in grades 2 through 11 were required to take the Stanford 9, a commercially available, nationally normed test. This satisfied Governor Wilson's desire to promote parental responsibility by sending home test scores for each child as soon as possible, but the Stanford 9 rapidly became the tail that wagged the accountability dog. As the accountability system continued to unfold under the next governor, Gray Davis, monetary incentives for teachers and schools were attached to test-score gains on the Stanford 9. Schools got the message and began preparing students for the material tested by the Stanford 9—even though they were supposed to be teaching the state's curriculum in order to prepare for the California-developed assessments that were to come later. Thus as usual, and not by design, the tests drove the standards and the curriculum rather than the other way around.

Foundations

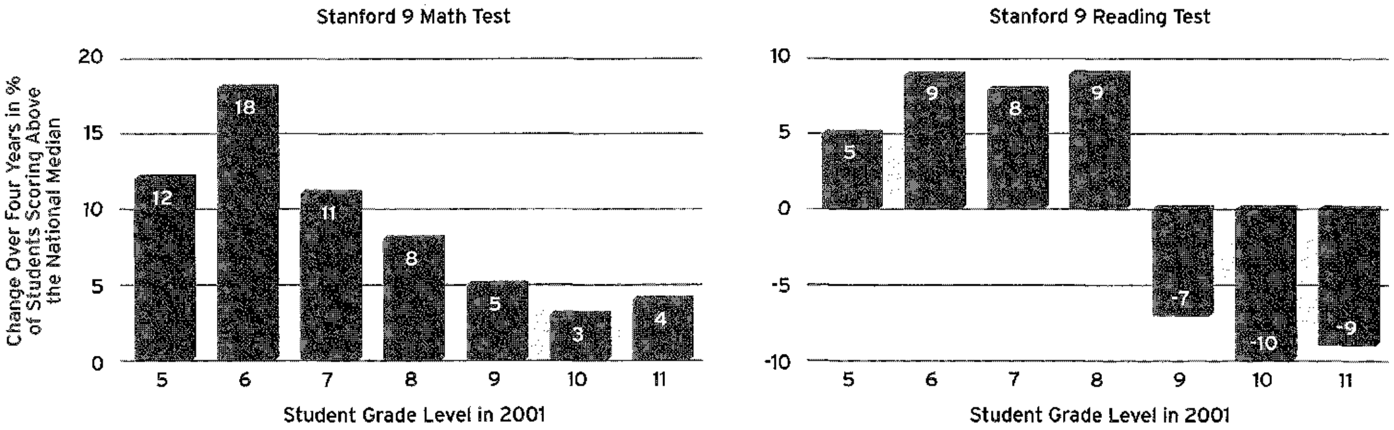
Davis, a Democrat, ran for governor in 1998 with education as his top priority, proclaiming, "Local control of California education is a disaster." His theory of change relied on rewards and sanctions for schools along with an informed public agitating for improvement. After Davis's election, he delivered on his promise to centralize accountability at the state level by persuading the legislature to pass legislation incorporating four basic elements:

- an academic performance index to measure each school based on its changes from year to year;
- an intervention program for underperforming schools;
- monetary awards for schools making gains and college scholarships for high-scoring students; and
- an exit exam that all students must pass to graduate from high school.

So far, the only component of the academic performance index has been the Stanford 9, since these were the only data available (see Figure 1). Each California school receives a ranking on the index from 1 to 10, and schools are also ranked against the schools in their income range. The release of the rankings

Mixed Results (Figure 1)

Younger students have made the biggest gains under California's STAR accountability system. The reading scores of older students have actually slipped.



SOURCE: California Department of Education's Education Data Partnership

Governor Gray Davis wants to spread the school-improvement grants among all schools that are below the average.

PHOTOGRAPH BY AP/WORLD WIDE PHOTOS

is a major media event and the rankings are published on the Internet. The state's standards-based assessment will be incorporated into the index this year. Other data, such as attendance and dropout rates, will hopefully be reliable enough to be added in the future.

The intervention program is reserved for schools that both fall below the Stanford 9 national average and do not meet their goal for gains on the Stanford 9. Test scores are disaggregated by race and ethnicity, but schools need only to raise their overall scores to avoid the intervention program. The state sets a goal of about a 5 percent gain for every school in the state. The growth target is set even higher if a school is grossly underperforming on the Stanford 9. As part of the intervention program, grants of $50,000 per school were first awarded to 304 schools during the 2000–01 school year. The funds were used to hire consultants who help plan and evaluate school improvement, followed by a $200 per-pupil implementation grant, to be used for anything from motivational programs for teachers to new instructional programs. The consultants are often retired teachers and school administrators, as well as staff from the U.S. Department of Education's regional education laboratories. Early evaluations of these plans have not been encouraging; many of the plans use strategies aimed at eliciting short-term test-score gains rather than inspiring long-term school improvement.

Schools that have failed to meet their growth target after two years of implementation are deemed low performing. This can trigger the state's intervention in various ways, including having the state superintendent take over a school. There are criteria for triggering a state intervention, but it is less clear how a school can get itself out of the doghouse. When is a school no longer judged to be low performing? There is also continuing disagreement over which schools should receive extra resources. Governor Davis wants to spread the school-improvement grants among all schools that are below the average—roughly 50 percent of the schools in the state—because he wants to build a centrist political coalition, but his fellow Democrats want to concentrate funds on the 20th percentile and below. The outcome will depend on the state's budget, however. State-led interventions may have to be postponed in a recession because of the need to cut funding, but the legislature is more willing to cut the rewards than the funds for state interventions.

In 2001 Governor Davis also provided $667 million in school-performance awards for test-score gains on the Stanford 9. In order to receive awards, schools must produce both overall gains and gains for various ethnic and racial groups. One program gives each teacher in a school $25,000 if the school greatly exceeds the state target for Stanford 9 increases. An additional $135 million was provided for college scholarships to 11th graders who scored high on the Stanford 9, despite concerns that the Stanford 9 tests mainly basic skills, while college-bound high schoolers are expected to take courses focusing on English literature and algebra.

The final element of the current accountability system is that by 2004 all students must pass a high-school exit exam that is based on the state standards for grades 7 through 10. This is more of a minimum-skills test than an exit exam, but it at least includes two essays.

Wilson and the Democrat-controlled legislature believed California schools were in such bad shape that they could not wait for the sensible sequential process of developing a distinctive California test based on California's curriculum standards. This was the equivalent of saying, "Fire, ready, aim." The controversy has not died out, even though scores on the Stanford 9 have been going up every year since 1998, and advocates are claiming victory. For example, Ron Unz, who sponsored a successful state initiative to restrict bilingual education, issued a press release saying that English immersion was a major cause of the rising scores. However, the state legislature's independent Legislative Analyst office, headed by Elizabeth Hill, warned that increases in scores could be the result of the fact that questions on the Stanford 9 remained similar for three years, so students and teachers are becoming more familiar with the style of questions.

By 2000, the entire state leadership realized that something had to be done to better align all the facets of the accountability system and to lessen the impact of Stanford 9 testing. The California-developed, standards-based assessments were first administered in 2001. These include secondary-school end-of-course exams that, like New York State's Regents exams, are arousing the interest of colleges and universities as they look for ways, besides the SAT and grades, to measure students' ability and motivation. Substantial funds have been appropriated for professional development and textbooks aligned with the standards-based assessments. Standards for teacher training were recently aligned with the state's curriculum and assessments. For now, a shorter form of the Stanford 9 will be used in most subjects, and it will have less impact on the incentives and sanctions that are being doled out. In three years, California may become a national leader of the accountability movement, as all the major components of a coherent and rational accountability system fall into place.

More Obstacles

Of course there is still much work to be done. The policymaking is largely over, but the campaign to change classroom instruction has just begun. Local educators are keenly aware of the state's accountability pressure, but awareness does not equal commitment to the state's goals or to classroom change. The accountability system must win the support of teachers, and schools and teachers need the training and resources necessary to help them teach the state's challenging standards. Unfortunately, California is running out of money just as its accountability system is coming together. The energy crisis drained the state of its surplus that was generated by the technology boom of 1995–2000, and the economy, dependent as it is on the fortunes of Silicon Valley, isn't what it was three years ago. The state surplus had helped to expand standards-based professional development programs, including algebra academies for teachers and students. But Governor Davis had to cut the teacher institutes in 2001, and the state is forecasting an $12.5 billion deficit in 2002. The governor proposed to cut the 2002 program of incentive payments to teachers and schools by 60 percent, and the legislature voted to cut it by even more. Meanwhile, differences across schools in per-pupil funding continue to create differences in learning opportunities, and this issue will only become more visible as the full accountability system is implemented and poor schools fail in large numbers. The state is again facing a lawsuit, this time brought by the American Civil Liberties Union, claiming that it is not providing an adequate education to all children.

There is also the challenge of moving past the strong signals that have been sent to educators that the Stanford 9 test *is* the accountability system. Moreover, local educators claim that there are still too many secondary-school tests, including four statewide assessments (the Stanford 9, the high-school exit exam, the standards-based exams, and the Golden State Exam, a holdover from the 1980s). California's colleges and universities add two admissions tests (SAT I and II); and the University of California, California State University, and the community colleges use three different placement exams. These assessments somehow need to be rationalized.

Policymakers must also deal with the inevitable backlash. Statewide, only 1 percent of students have opted out of testing, but a very vocal set of parents in San Diego and Marin County have refused to let their children take the state test. These represent only a fraction of the parents in high-income suburbs who believe that standards-based accountability will only lead to more state interference, which will push the teaching in their school to the lowest common statewide level. Wayne Johnson, president of the California State Teachers Association, has called the high-school exit exam a "disaster." He predicts that students whose English is limited will fail in large numbers and claims that the math questions are too difficult. A few teachers rejected their Stanford 9 incentive bonuses, calling them "blood money" and "bribery." Earlier concessions to the teacher unions forbade the use of state tests for hiring, firing, or promoting teachers. Students' grades are based primarily on teacher-designed or local tests, so the stakes for students are unclear.

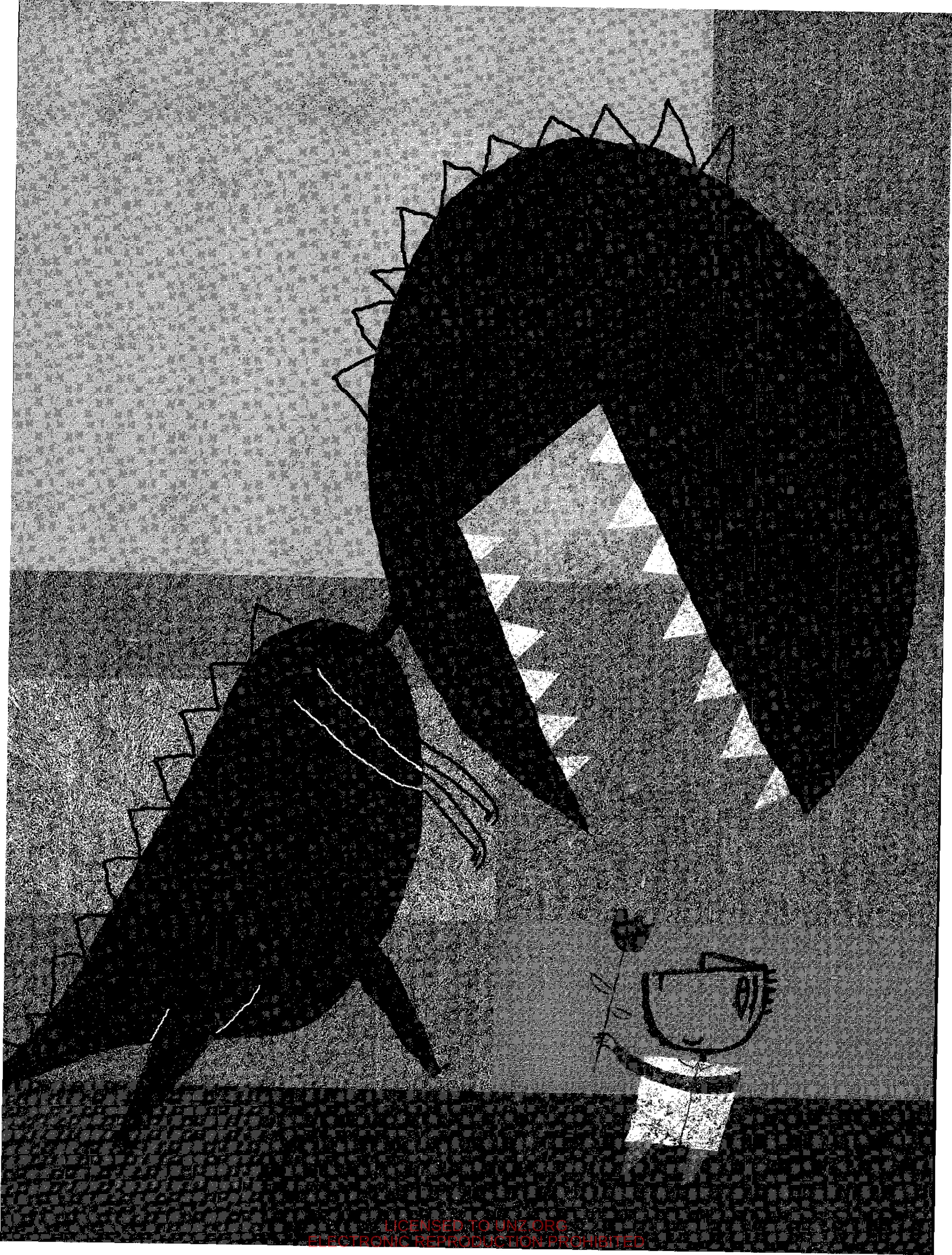
Adding to the confusion is the new federal education law that President Bush signed in 2002. The major problem is a federal requirement that by the fall of 2002 all teachers hired under Title I must be "highly qualified," and by 2005 every public school teacher must be "highly qualified." In 2001, 42,427 teachers (one out of seven statewide) were working in California without a preliminary credential that the state has defined as a minimum requirement. Several California districts hire teachers who have just a bachelor's degree and a passing score on a minimum skills test that is set at the 10th grade level. California officials have no strategy to fulfill this federal teacher quality requirement.

While the spirit of California's accountability system is in accordance with the federal law, a significant issue is the definition of pupil "proficiency." In 2001, California established five performance levels on its state test and defined "proficiency" as the level of achievement necessary to enter a university. Only 30 percent of California students are at the proficient level, but the federal law requires all students to be "proficient" in 12 years. California must either lower its standards for proficiency or negotiate a federal waiver. Moreover, California's accountability law measures annual pupil growth on a schoolwide basis, averaging all student scores in a school, while the federal law requires each student to meet growth targets.

California's experience with school report cards also raises questions about the federal proposals. The 2002 federal law requires that states provide annual report cards with a wide range of information for each district and school. California uses both an elaborate report card (with 50 criteria) and a school ranking (on a scale of 1 to 10) based solely on test scores. The long report card has the virtue of being comprehensive, but it is not used by educators and is too complex for the public. The ranking, by contrast, is easily grasped by parents and is noticed. A single number, however, oversimplifies the complexity of schooling. Parents should receive a few clear measures and an indication of where to get more information about their child's school. Overall, report cards don't cause much change, and the more data they contain, the more confusing they are to parents.

In the end, perhaps the instability of earlier accountability schemes will help to keep this one in place. Policymakers keep saying that California must stick with something, and the business community has formed a coalition to preserve the current agenda and direction. The real political test will come when students begin to fail the graduation exam, or when schools face dramatic interventions like faculty firings or reconstitutions. The key issue is whether the state's accountability system will change and improve classroom instruction. So far, the linkage between state policy and the classroom has been uncertain and largely indirect. Many state policymakers realize this and hope that a coordinated system will send clear signals to local educators. But it is unclear whether the current blend of exhortation, sanctions, incentives, and publicity will affect classroom instruction in a major way.

–Michael W. Kirst is a professor of education at Stanford University and director of Policy Analysis for California Education, a joint venture of the Stanford and University of California–Berkeley schools of education.



MONSTER HYPE

How a few isolated tragedies—and their supposed causes—were turned into a national "epidemic"

by JOEL BEST

CONTEMPORARY DISCUSSIONS ABOUT SOCIAL issues, especially within education, almost always involve statistics. Numbers have become an essential element in policy rhetoric, a form of evidence needed to persuade others. Statistics let us claim that we can measure the size of our problems and the effectiveness of our solutions.

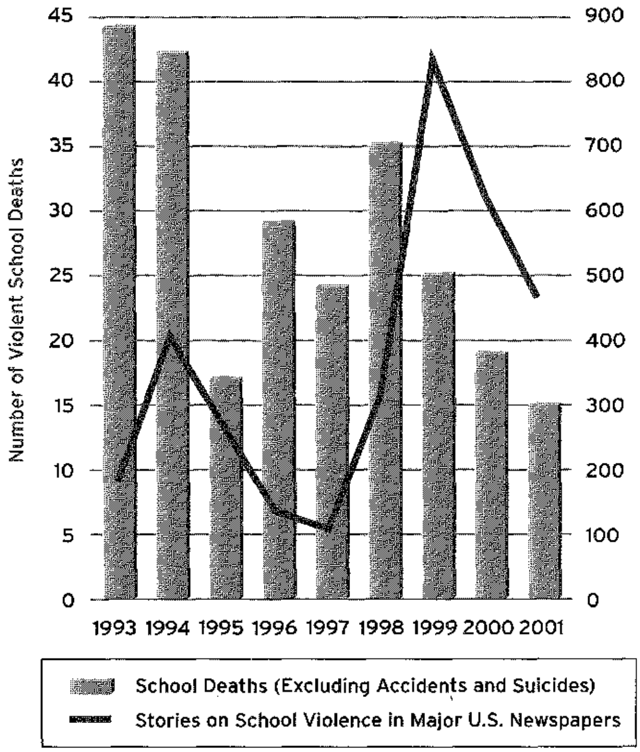
Yet even as we rely on numbers, we are bedeviled by innumeracy, the mathematical equivalent of illiteracy. Too often, we fail to think critically about the statistics we encounter, to ask even the most basic questions. This is important, because accepting numbers uncritically may cause us to badly misunderstand our problems. There are few better examples of this failing than some of the recent figures regarding school violence.

The March 5, 2001, shooting spree at Santana High School in Santee, California, which left 2 dead and 13 injured, revived concerns over the seeming escalation of school violence and its potential links to the age-old schoolyard tradition of bullying. School shootings first became a serious issue in the wake of a series of tragic incidents, the most famous being Dylan Klebold and Eric Harris's April 1999 rampage at Columbine High School, in Littleton, Colorado, during which they murdered 12 students and a teacher before turning their weapons on themselves. Of particular interest was that nearly every shooting was accompanied by reports that the teenagers involved were marginalized in some way; the Santee shooter especially appears to have been a victim of bullying.

ILLUSTRATIONS BY JAMES YANG

A Phantom Trend (Figure 1)

In the wake of the Columbine shootings, the media's increased coverage has created the false impression that school violence is on the rise. The number of news stories on school violence remains dramatically above the levels of the early '90s despite a continued decrease in the number of violent school deaths.



SOURCES: National School Safety Center; Lexis-Nexis

Both school shootings and bullying have become subjects of extensive media coverage, featuring the pontification of assorted politicians, activists, and experts. This is how contemporary Americans create new social problems. Typically, the process involves a three-part recipe:

1) Illustrate the problem with an awful example (e.g., the mass murder at Columbine High School).
2) Give the problem a name ("school shootings").
3) Use statistics to suggest the problem's size and importance.

Statistics play a crucial role in this process, because we tend to assume that numbers are factual—that somebody has counted something, that the problem has been measured and therefore is as big as the claims suggest. Coupled with dramatic, headline-grabbing incidents, they have created the impression that both school violence and bullying are on the rise. This may make for compelling television, but the oversaturated media coverage can portray a few isolated incidents as a national trend. Take CBS anchor Dan Rather's post-Santee warning: "School shootings in this country have become an epidemic." Such claims have become commonplace among journalists who haven't thought carefully enough about the evidence. The statistics on violence and bullying that are trucked out by pundits and activists often exaggerate or distort the case. The result is that the public and policymakers tend to overreact as they look for solutions to problems that appear to be out of control. A closer look at the statistics, however, reveals a more complicated and hopeful picture.

A Phantom Epidemic

Of course, the phenomenon of adolescents' bringing guns to school and randomly shooting their peers ought to be a source of genuine worry. The 1997–98 school year alone saw tragedy strike West Paducah, Kentucky (3 dead, 5 more wounded); Jonesboro, Arkansas (5 dead, 10 wounded); and Springfield, Oregon (2 deaths and 21 wounded at the school, after the shooter first killed his parents at home). These crimes, along with the Columbine massacre, seemed to be without rational motivation; what could possibly have driven these adolescents to lash out at the world in such bloody fashion? Kids have always divided into cliques and subjected the nonconforming to verbal and physical abuse; only recently, it seemed, had the social drama of high school resulted in mass casualties. It's legitimate to wonder whether these incidents represent a deep-seated change in youth culture.

Nevertheless, these tragic events masked the overall trend: a good deal of evidence indicates that school violence has actually been declining in recent years. When researchers at the National School Safety Center (NSSC) combed media reports from the school years 1992–93 to 2000–01, they identified 321 violent deaths at school. However, not all of these incidents involved student-on-student violence: they included, for example, 16 accidental deaths and 56 suicides, as well as incidents involving nonstudents, such as a teacher killed by her estranged husband, who then shot himself, and a nonstudent killed on a school playground during the weekend. Even if we include all 321 deaths, the average fell from 48 violent deaths per year during the school years 1992–93 through 1996–97 to 32 per year from 1997–98 to 2000–01. If accidental deaths and suicides are eliminated from the data, the decline remains: from an average of 31 deaths per year in the earlier period to 24 per year in the later one. Moreover, the later period includes all of the heavily publicized cases mentioned above. And the later figure may be further inflated by the likelihood that the media were more apt to report school shootings after the topic vaulted to public attention (see Figure 1).

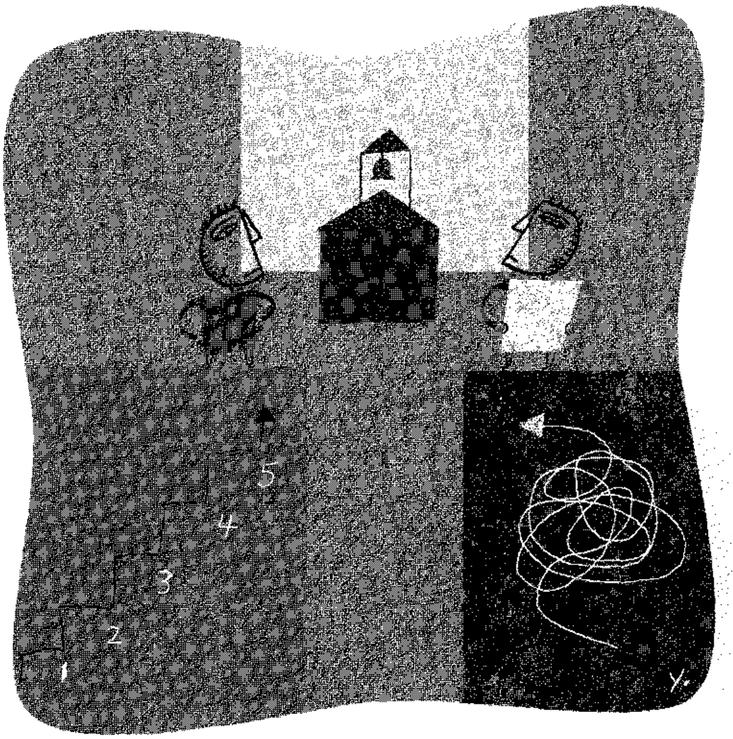
This decline is consistent with the evidence suggesting that crime rates were declining nationwide. During the 1990s, the overall crime rate fell, as did the rates of major violent crimes such as homicide, robbery, and aggravated assault. The crime rate, which is the Federal Bureau of Investigation's tally of crimes reported to the police, is only one of two national measures of criminal activity. The second, less familiar measure is the rate of victimization reported in the National Crime Victimization Survey. Researchers with the victimization survey interview a large national sample and ask respondents whether they or anyone in their households have been victims of crime. This survey showed instances of criminal victimization falling during the 1990s. Moreover, reports of teenagers being victimized by violent crimes at school dropped. The data also showed that instances of victimization were less common at school than elsewhere; in other words, teenagers were safer at school.

The federal Centers for Disease Control and Prevention's Youth Risk Behavior Survey also found steadily declining percentages of high-school students who reported fighting or carrying weapons on school property during the 1990s. It is also important to recognize that the risks of school violence are extremely low. For every million children who attend school, there is less than one violent school-related death per year. Moreover, only about 1 percent of children killed by violence are hurt at school, despite the large amount of time they spend there.

None of these data are especially hard to come by; all of them were readily available—and the trends they showed were apparent—before, during, and after the various school-shooting incidents that became subjects of extensive news coverage. All of this evidence flatly contradicted the claims that there was a wave, trend, or epidemic of school violence. In other words, the wave of school shootings was a phantom—that is, a nonexistent trend. What accounts for this misperception? Why did the press and the public assume that school shootings were increasing?

In large part, media coverage promoted this distorted view of the problem. The Columbine killings in particular became a huge story. Columbine involved many victims, and the story unfolded over hours. Because the crime occurred in the suburbs of a major city, there were plenty of reporters nearby, and they had time to arrive on the scene for live coverage. The result was dramatic video footage that would be replayed many times. Furthermore, Columbine was a bastion of suburban privilege; it challenged stereotypes about inner-city violence. It was a story made for television.

The Columbine coverage also reflected recent media transformations. Most Americans now have access to cable or satellite television systems; they are no longer limited to receiving broadcasts from a handful of local stations. Most viewers now can choose among several all-news or public-affairs channels. Those channels need constantly to fill the time with content. In the aftermath of the Columbine shootings, broadcasters like CNN, Fox News, and MSNBC devoted hours, not just to reporting the story and commentary about the violence, but also to live coverage of many funeral and memorial services. Columbine remained a major story for days, and during that period, politicians, activists, and commentators used it as evidence to justify their calls for a wide range of measures, including tougher gun laws, restrictions on adolescents' access to violent popular culture, and so on.

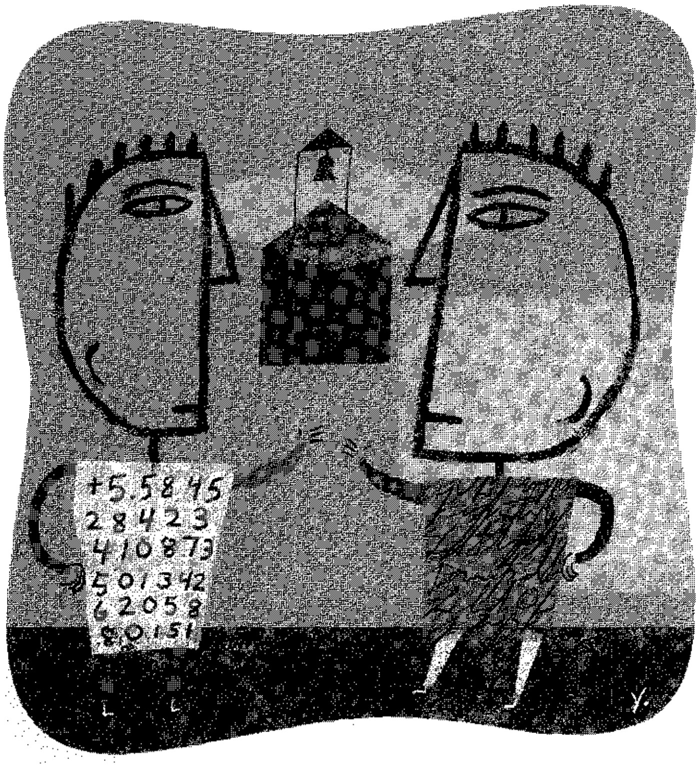


Coupled with dramatic, headline-grabbing incidents, the misuse and abuse of statistics have created the impression that both school violence and bullying are on the rise.

The Columbine killings were a terrible event, but we are accustomed to thinking about such incidents as instances—that is, as examples of some larger problem. The extraordinary level of media coverage reinforced the interpretation that these killings must have had some larger significance. It also gave people the sense that school shootings must be a large and growing problem, regardless of what the available statistics actually showed.

Making Bullies Seem Big

Concern with school bullying usually concentrates on mundane,



These tragic events masked the overall trend: school violence has actually been declining in recent years.

everyday cruelty, but the issue has taken on new significance with reports that several school shooters were reacting to a history of taunts and shoves from their peers. All of us have witnessed—if not experienced—bullying, usually during our own school days. The phenomenon has long been recognized as undermining efforts to maintain discipline in schools.

The problem with trying to measure the extent of schoolyard bullying should be obvious: the term lacks any clear definition. What is bullying? Any attempt to measure bullying requires not just defining bullying in the abstract, but also devising an operational definition—that is, a set of procedures or operations whereby one can identify and count cases of bullying. In general, those who are trying to provoke a policy response to a particular social problem favor broad, inclusive definitions. Their purpose is to make the problem seem as widespread as possible. If bullying is defined narrowly—say, as involving only physical assaults—it will seem like less of a problem than if the definition includes many other forms of hurtful behavior.

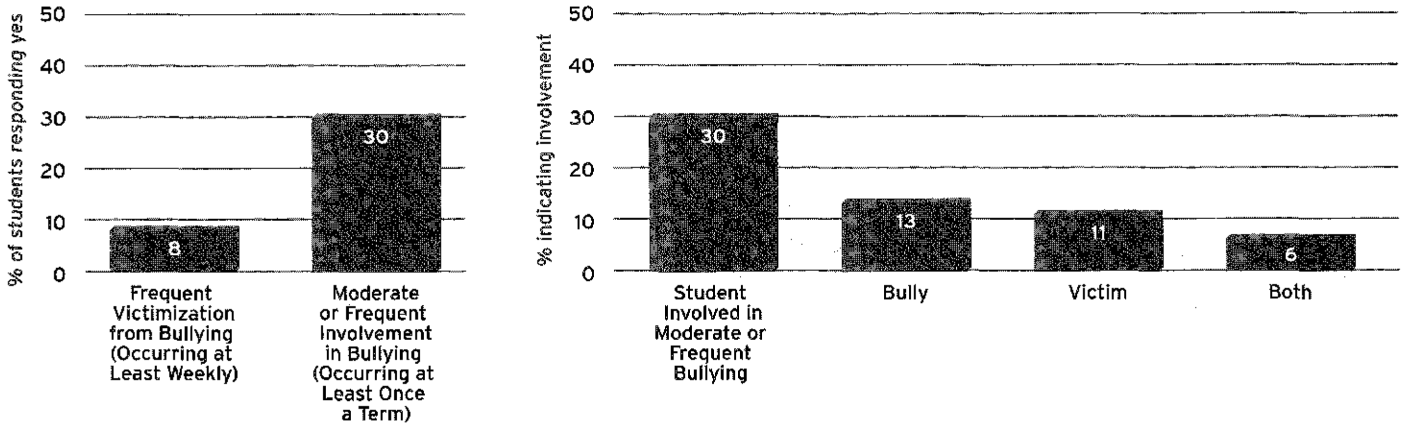
Consider "Bullying Behaviors Among U.S. Youth," an article published in the April 25, 2001, issue of the *Journal of the American Medical Association (JAMA)*. The article reported results from a large (nearly 16,000 respondents) representative sample of students in grades 6 through 10; the authors were associated with the National Institute of Child Health and Human Development (which supported the survey). Its major finding, widely reported by the media, was that nearly 30 percent of youths "reported moderate or frequent involvement in bullying." This was just one of many studies of bullying, but few other studies feature samples as large and well drawn. In addition, this article appeared in *JAMA*, an especially prestigious journal, so we might assume that it represents the best work on bullying. Nevertheless, it is hard to imagine JAMA finding the space to publish these results if the researchers had found that only 3 percent of youths were involved in bullying, let alone to picture the news media heralding the article's publication.

Of course, the proportion of surveyed students who report being involved in bullying will depend on how bullying is defined and how the questions are framed. The questionnaire's section on bullying began with an explanation: "Here are some questions about bullying. We say a student is BEING BULLIED when another student, or a group of students, say or do nasty and unpleasant things to him or her. It is also bullying when a student is teased repeatedly in a way he or she doesn't like. But it is NOT BULLYING when two students of about the same strength quarrel or fight." The students were then asked how frequently they bullied others or were bullied during the current school term. They had to choose from the following responses: "I haven't . . . ," "once or twice," "sometimes," "about once a week," or "several times a week." In presenting their results, the authors defined incidents that occurred at least weekly as frequent involvement in bullying, and responses of "sometimes" as moderate involvement.

The authors conclude that "bullying is a serious problem for U.S. youth," and that "the prevalence of bullying observed in this study suggests the importance of preventive intervention research targeting bullying behaviors." But the data are much less clear than these conclusions suggest. The central finding—that 30 percent of youths are involved in bullying—depends on three manipulations, three methodological choices. First, students could be involved either as a bully (13 percent), as a victim (11 percent), or as both (6 percent) (see Figure 2). If the article's authors had chosen to count only the victims, they would have found that 17 percent, rather than 30 percent, were subject to bullying. Second, the authors included both "moderate" bullying (those who reported bullying as occurring "sometimes"—more than once or twice during the term, but less than weekly), and "frequent" bullying (occurring at least weekly). Adopting a narrower definition would have made the findings less dramatic; only 8 percent of the respondents reported being targets of frequent bullying. Third, there were separate questions about bullying in and out of school, but those responses were combined for the *JAMA* article, so readers are not told how much of the reported bullying actually

How Serious a Problem Is Bullying? (Figure 2)

The media widely reported the finding that 30 percent of students said they were involved in moderate to frequent bullying. But that number falls to just 8 percent if only the students reporting frequent victimization from bullying are included. Moreover, the 30 percent figure falls to only 11 percent if limited to victims who say they have not also been bullies.



SOURCE: "Bullying Behaviors Among U.S. Youth," *Journal of the American Medical Association*, 2001

occurred in schools. We can assume that at least some of the students who reported being frequently bullied were targeted outside of school. If so, the percentage of students who reported frequent bullying in school should have been even lower than 8 percent.

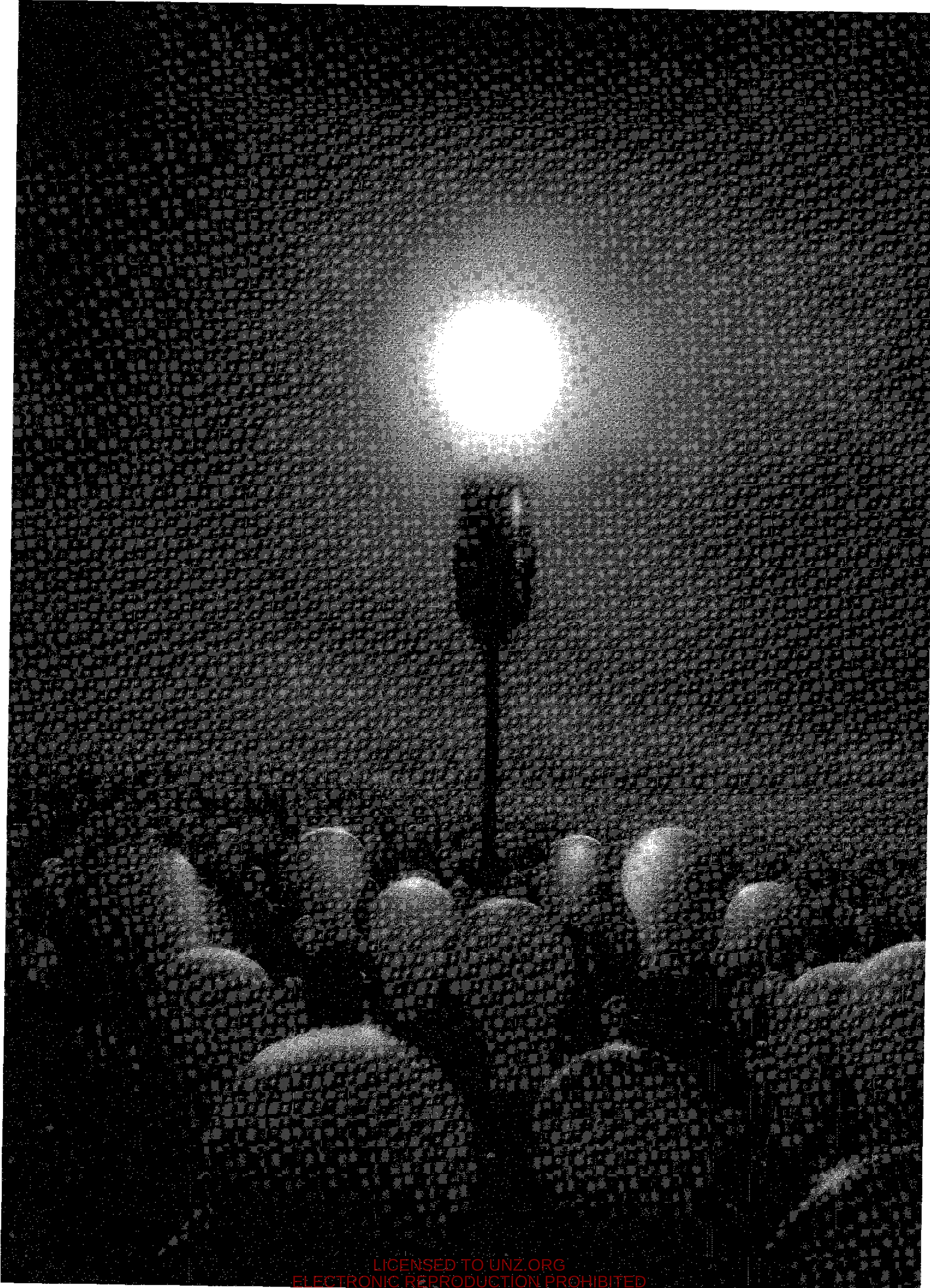
In other words, the authors made a series of choices that allowed them to estimate that bullying significantly affects 30 percent of students; different choices—say, looking only at victims of frequent bullying in schools—would have produced a figure that was perhaps a quarter as large. The point is not that this is a bad piece of research, nor that bullying can't have serious consequences (remember the Santee shooter). Rather, it is that the numbers that emerge from social-science research need to be handled with some care. How the survey's questions were worded, the order in which they were asked, and the choices made in interpreting and summarizing the data for publication all shaped the results.

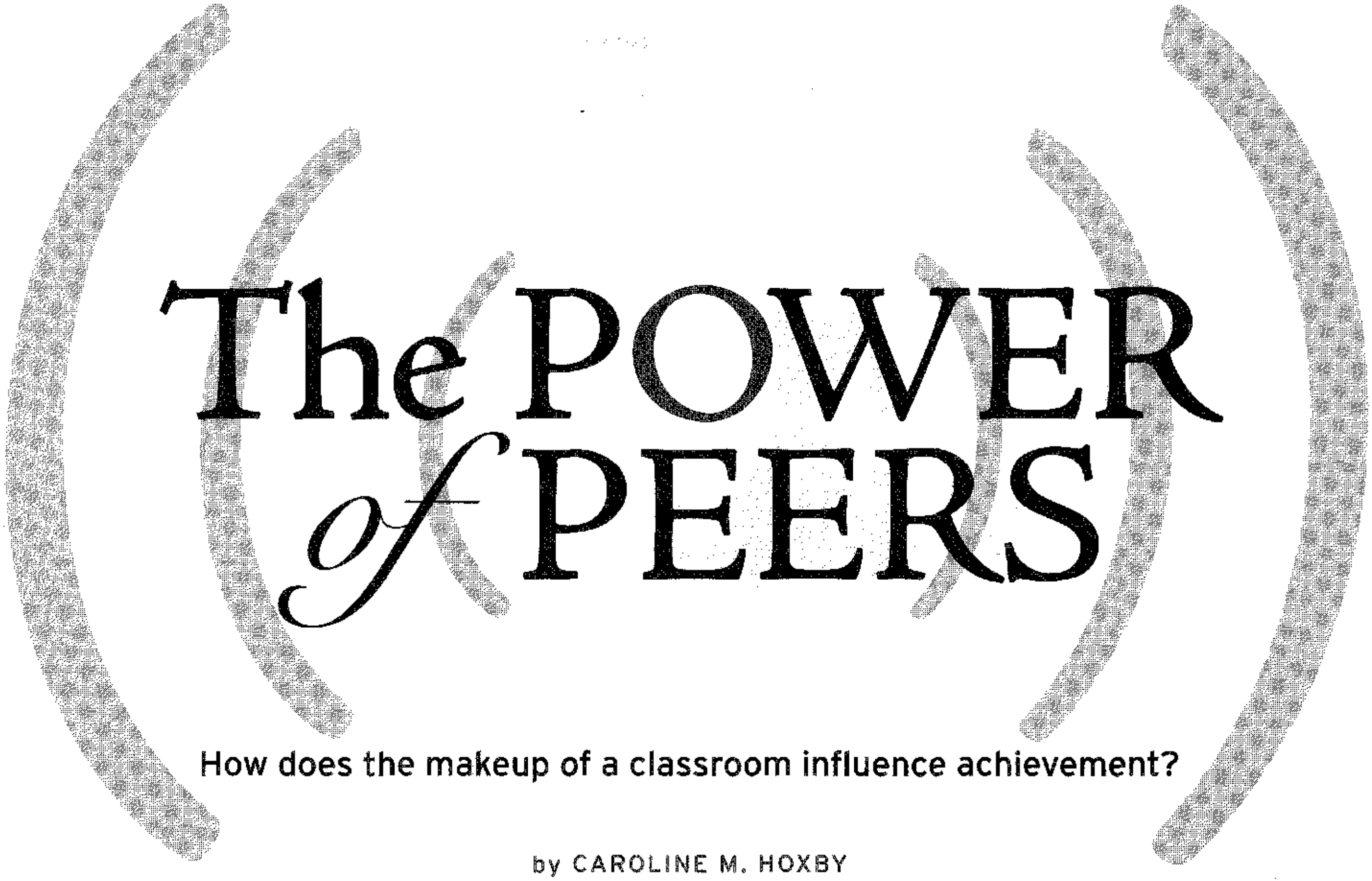
Most of us depend on statistics; we need them to summarize our world, which is otherwise too large and complex to grasp. We equate numbers with facts, and figures have become an essential element in contemporary policy discussions. They are used to document the existence, size, and other dimensions of social problems. Still, even though we depend upon statistics, we need to view them critically. Statistics are not little nuggets of truth. Rather, they are products of someone's efforts. Any number means that someone has gone to the trouble of counting something. This raises questions: Who did the counting, and why? What did they count? How did they go about counting it? Why did they choose to count what they did, and to go about it in the way they did?

Perceptions that school violence is a national crisis have clearly affected educational policy. Resources have been directed toward purchasing and operating metal detectors, stationing police officers in schools, and enhancing other security measures, restrictions that many schools expanded in the aftermath of the September 11 terrorist attacks (see Education Matters to Me, page 88). These are not necessarily bad choices, but they may have been based in part on inaccurate perceptions of the nature and level of school violence.

> The authors made a series of choices that allowed them to estimate that bullying significantly affects 30 percent of students. Different choices would have produced a number that was a quarter as large.

–Joel Best is a professor of sociology and criminal justice at the University of Delaware. His latest book is Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists *(University of California Press, 2001).*





The POWER of PEERS

How does the makeup of a classroom influence achievement?

by CAROLINE M. HOXBY

What does the term "peer effects" mean in a school environment? It includes the effects of students' teaching one another, but that is only the most direct form of peer effects. Intelligent, hard-working students can affect their peers through knowledge spillovers and through their influence on academic and disciplinary standards in the classroom. Alternatively, misbehaving students may disrupt the classroom, thereby sapping their teacher's time and energy. The makeup of a classroom—its average family income, the number of children with disabilities, its racial and gender balance—can also create peer effects. Children with learning disabilities may draw disproportionately on their teacher's time; racial or gender tension in the classroom may interfere with learning; wealthier parents may purchase learning resources that get spread over a classroom. Peer effects may even operate through the ways in which teachers or administrators react to students. For instance, if teachers believe that less should be expected of minority children, they might lower their academic standards when confronted with a classroom that has a high share of black or Hispanic students. The other students in such a classroom would experience negative peer effects, not due to the minority students' influence but because of the teacher's assumptions.

Peer effects, if they do indeed exist, have implications for a number of policy issues in education. For example, the literature on school finance and control is currently absorbed with the question of whether students are affected by the achievement of their schoolmates. If peer effects exist at school, a school-finance system that encourages an efficient distribution of peers among all schools will make society's investments in student learning more productive. The debate over "tracking," the system in which students are exposed only to peers with similar achievement, turns partly on the

ILLUSTRATION BY TINA WEST/WWW.IMAGES.COM

PHOTOGRAPH BY HULTON/ARCHIVE

question of whether being concentrated in lower-level classrooms merely exacerbates the problems of low-achieving children. Desegregation plans that assign students to schools outside their neighborhood or school district also rest partly on the belief that one's peers can exercise enormous influence over one's performance.

However, there are two principal difficulties for theories that rely on peer effects. First, however much sense the theory of peer effects makes, there are formidable obstacles to estimating them. Although some credible estimates of peer effects do exist, people often rely on evidence that is seriously biased by selection effects. For instance, if everyone in a group is high achieving, many observers assume that achievement is an *effect* of belonging to the group instead of a *reason* for belonging to it.

Also, the most popular model used by researchers to estimate peer effects (the "baseline" model) assumes that peer effects are a zero-sum phenomenon—that is, in order to give one student a better peer, that peer must be taken away from another student; the two effects cancel one another out. According to the baseline model, a student's reading score would be affected linearly by the average reading score of his classmates. Regardless of how one allocates peers, total societal achievement remains the same under the baseline model. But many arguments assume that total societal achievement can be increased if peers are redistributed. For instance, the argument against "tracking" is based on the notion that both low- *and* high-achieving students benefit from being exposed to one another in the classroom. By contrast, the idea behind "gifted and talented" programs is that high-achieving students benefit from being among one another. Thus, although it is tempting to dismiss the baseline model as naive or restrictive, if one were able to show, empirically, that the baseline model adequately described peer effects, some interesting theories would fall by the wayside.

The central problem with estimating peer effects in schools is that families, in a number of ways, can select their children's peers. Families self-select into schools based on their incomes, job locations, residential preferences, and educational preferences. A family may even self-select into a school based on the ability of an individual child; a family with a highly able child may choose to live near a school that has a program for gifted children. Moreover, families may influence the particular classes to which their children are assigned within their schools. If, for example, savvy parents believe that a certain 3rd grade teacher is particularly good, they may get their children assigned to her class, thereby creating a classroom of children whose parents care about education to an unusual degree. School administrators and teachers can also select students into particular classrooms for reasons that are related to achievement. For instance, a school may assign children with similar achievement levels to the same classroom, in order to minimize teaching difficulty. Or a school may place all of the "problem" students in a certain teacher's class because she is good at dealing with them. In short, it should be assumed that a child's being in a certain school, and even a particular classroom, *is* associated with unobserved variables—such as highly involved parents—that affect his achievement.

New Strategies

This study introduces two empirical strategies that circumvent these obstacles by examining differences in *cohorts* of students—a school's group of 3rd graders in one year versus the next year's group of 3rd graders—rather than cross-sections of classrooms at the same grade level. Both strategies depend on the idea that the peer composition of a certain grade within a school—its gender and racial balance, its mix of the studious and the troublesome—will vary from year to year in a way that is idiosyncratic and beyond the easy management of parents and schools. Even within a school that has an entirely stable population of families, timing and simple biological variation would create birth cohorts that idiosyncratically vary in their natural talents and racial and gender makeup.

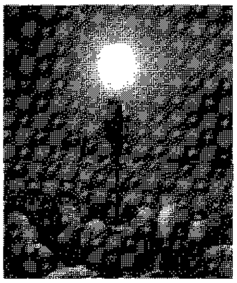
Suppose, for instance, that a family shows up for kindergarten with their older son and finds that, simply because of random variation in local births, their son's cohort is 80 percent female. The next year, they show up with their younger son and find that, also because of random variation, his cohort is 30 percent female. Their older son will be exposed to more female students (who tend to be higher achievers in elementary school). Their younger son will be exposed to more male students. Because the two boys have the same parents and the same school, the main difference in their experience will be their peers.

It is this type of unexpected variation in cohort composition that the empirical strategies in this study attempt to exploit. A parent may have a fairly accurate impression of the cohorts around his child's age, and may pick a school on that basis, but it is difficult for a parent to react to a cohort composition "surprise" by changing schools. As long as we focus on idiosyncratic variation in cohort composition, as opposed to classroom composition, we need not worry about how schools and parents manipulate the assignment of students to classrooms. If a cohort is more female than the previous cohort, for instance, the school must allocate the "extra" females among its classrooms somehow. Inevitably, some students in the cohort will end up with a peer group that is more female than is typical.

In the first strategy, I attempt to identify idiosyncratic variation in cohort composition by comparing adjacent cohorts' gender and racial makeup. I see whether differences in the achievement of adjacent cohorts within a certain grade within a certain school are systematically associated with differences in the gender composition of those cohorts. If there are no peer effects, the average achievement of male (or female) students should not be affected by the *share* of their peers who are female.

In the second strategy, I attempt to identify the idiosyncratic component of each gender and racial group's achievement and determine whether the components are related to one another. For instance, if the females in the 1996–97 cohort of 3rd graders in School I have unusually low achievement, does one find that the males in the 1996–97 cohort of 3rd graders in School I have unusually low achievement too? If the Hispanic students in the 1994–95 cohort of 5th graders in School II have unusually high achievement, does one find that the white, black, and Asian students in the 1994–95 cohort of 5th graders in School II have unusually high achievement too?

This strategy requires an unbiased estimate of the idiosyncratic component of each group's achievement that is *independent* of the estimates with which one plans to correlate it—that is, the idiosyncratic component of the average achievement of other groups within the same cohort. This is accomplished by using only that portion of the achievement of a gender or racial group that cannot be explained by a linear time trend and the overall gender and racial composition of the group's cohort.



Both boys and girls tend to perform better in reading and math when they are in classes with larger shares of girls.

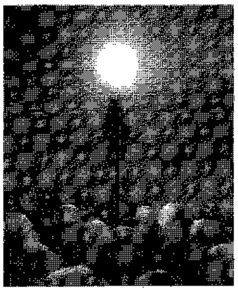
For both strategies, I am sensitive to the potential criticism that what appears to be idiosyncratic variation in racial groups' shares or achievement may actually be a time trend within a grade within a school. To address this criticism, I eliminate not only linear time trends but also any school in which actual years explain more variation (in cohort composition or in achievement) than false, randomly assigned years. These empirical strategies are, I would argue, an improvement on previous methods of identifying peer effects in schools. Previous researchers have most often estimated models like the baseline model and used cross-sectional variation in schoolmates to identify peer effects. They have dealt with the problem of selection bias by controlling for observable variables, comparing the educational experiences of siblings in families that have moved (so that the siblings experience different schools), studying children in magnet or desegregation programs, or estimating a selection model. In particular, Boston's METCO program, in which inner-city minority children are sent to schools in the suburbs, has been much studied. The difficulty with estimates based on programs like METCO is that children who enter the program (and do not leave it) are likely to have higher unobserved ability or motivation. In practice, these methods have generally proved unconvincing because there are unobservable variables that are correlated with peer selection, with moving, with participating in a magnet or other school program, or with the excluded variables that identify the selection model.

Only in some cases am I able to distinguish among the various channels through which peer effects can operate. In general, the peer effects estimated in this study (and in most research) embody multiple channels. In judging the magnitude of the results, it is important to keep the multiple channels in mind.

Data

Data for the entire population of 3rd, 4th, 5th, and 6th graders in public schools in the state of Texas during the 1990s were used. Beginning with the 1990–91 school year, Texas began to administer a statewide achievement test, the Texas Assessment of Academic Skills (TAAS), to elementary-school students. Scores on the TAAS form the basis of the analysis. Texas contains a very large number of elementary schools, which is fortunate because idiosyncratic variation in the gender and racial makeup of cohorts within a grade within a school is sufficiently uncommon that a large number of observations are needed to generate the necessary number of "natural events."

In a typical year during the 1990–91 to 1998–99 period, there were about 3,300 schools in Texas that enrolled 3rd graders; the size of the median cohort was about 80 students. Third graders were typically 49 percent female, 0.3 percent Native American, 2 percent Asian, 15 percent black, 33 percent Hispanic, and 49 percent Anglo. There were no apparent time trends in the shares of 3rd graders who were female or Native American. There were slight upward trends in the shares of third graders who were Asian (2.2 to 2.5 percent over the period), black (14.8 to 15.7 percent over the period), and Hispanic (30.7 to 34.9 percent). There was a mild downward trend in the share of 3rd graders who were Anglo (52.2 to 46.4 percent). The statistics for grades 4, 5, and 6 were very similar (naturally, because most of the students remain in these schools from year to year).



Young girls tend to be better readers than young boys. On the 3rd grade reading test, the average female scored 1.1 points—about half a standard deviation—higher than the average male. Some ethnic differences were even larger. Compared with the average white student, the average black student scored 3.6 points lower; the average Hispanic student, 2.9 points lower; the average Asian student, 0.7 points higher; and the average Native American student, 1.5 points lower. The black-white and Hispanic-white score gaps are substantial: 1.6 and 1.3 standard deviations, respectively.

There was an upward trend in the reading scores of all groups over the period, the average score rising from 28.5 to 31.3 points. Some improvement typically occurs during the first few years of administering a new test, simply owing to comfort with the test. The improvement in Texas accelerated over time, however, and the past few years' improvement are most likely due to true learning of the material tested by the examinations—particularly as Texas's curriculum and tests became more closely aligned.

There was a slight upward trend in math scores as well: an average gain of 0.1 points per year. The average female scored 0.1 points higher than the average male—a difference of only 0.03 standard deviations. Compared with the average white student, the average black student scored 4.7 points lower; the average Hispanic student, 3.2 points lower; the average Asian student, 1.3 points higher; and the average Native American student, 1.9 points lower. The black-white and Hispanic-white score gaps are again substantial: 1.6 and 1.1 standard deviations, respectively. The results on the 4th, 5th, and 6th grade tests were very similar to those in 3rd grade.

Results of Strategy 1

Gender. Both boys and girls tend to perform better in reading when they are in classes with larger shares of girls (see Figure 1). For instance, in 3rd grade reading, girls' scores rise by 0.037 points for every 10 percentage point change in the share of their class that is female. Males' scores rise by 0.047 points for every 10 percentage point change in the share of their class that is female. To put this in perspective, an all-female class would score about one-fifth of a standard deviation higher in reading, all else being equal. The effects for 4th, 5th, and 6th grade reading scores are similar. A translation of the results in a way that reveals the effects of peer achievement provides a different perspective: being surrounded by peers who score 1 point higher on average raises a student's own score by 0.3 to 0.5 points, depending on the grade. The translation suggests that peer effects are substantial.

Boys and girls also perform better in math when

Perhaps in female-dominated classrooms, girls do not experience pressure to be feminine and therefore remain enthusiastic about math.

they are in classes with larger shares of girls. In 3rd grade reading, girls' scores rise by 0.038 points for every 10 percentage point change in the share of their class that is female. The effect is larger in higher grades: female 6th graders' scores rise by 0.064 points for every 10 percentage point change in the share of their class that is female. Likewise, male 3rd graders score 0.040 points higher and male 6th graders score 0.081 points higher for every 10 percentage point change in the share of their class that is female. Because the average female scores only a little higher than the average male, however, the earlier translation of the scores generates implausibly large effects. If the translated effects were taken literally, one would conclude that being surrounded by peers whose math scores were on average 1 point higher would raise a student's own score by 1.7 to 6.8 points, depending on the grade. These effects are so large that they suggest that peer effects do not operate purely through the channel of peers' achievement in math.

There are a few alternative channels that might explain the effect of females on math scores. First, since learning math requires reading, and reading scores are higher in classes with higher percentages of females, females may affect subjects like math *through* their (quite plausible) peer effect on reading. Second, classes with more girls may simply have fewer disruptive students or a more learning-oriented culture. Third, classroom observers have argued that the pressure to be feminine makes girls unenthusiastic about math. Perhaps in female-dominated classrooms, girls do not experience this kind of pressure and therefore remain enthusiastic about math—thereby allowing the teacher to teach it better to all students. In any case, it is clear that the baseline model of peer effects is inadequate: peer effects do not operate solely through peers' mean achievement in the same subject.

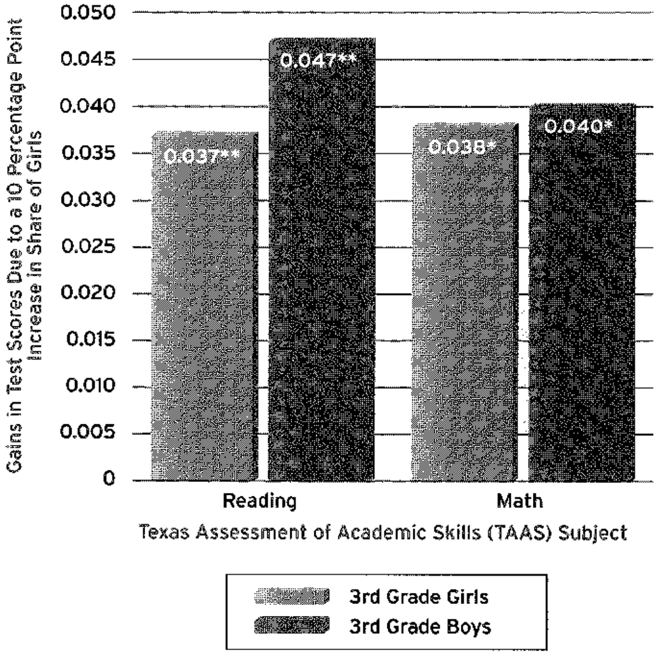
Race. In interpreting the next set of results, it is worthwhile to remember that the peer effects of any racial group include the effect of variables associated with that group, including their family income, parents' education and level of involvement, and the language spoken in the home. They should not be interpreted as the effects of a group's innate ability. In particular, black and Hispanic students are far more likely to be poor than are white students in Texas. Therefore, any negative peer effects associated with being in classes with large shares of black students largely reflect the impact of being exposed to low-income students.

Black, Hispanic, and white 3rd graders all tend to perform worse in reading and math when they are in classes that have a larger share of black students. For every 10 percentage point rise in the share of their class that is black, black students' reading scores fall by 0.250 points, Hispanic students' reading scores fall by 0.098 points, and white students' reading scores fall by 0.062 points. For the same 10 percentage point change in the share of their class that is black, black students' math scores fall by 0.186 points, Hispanic students' math scores fall by 0.086 points, and white students' reading scores fall by 0.043 points. What's particularly interesting is that having more black peers appears to be most damaging to other black students. Recalling that black students have the lowest scores on both the reading and math tests, one can see that these results can be interpreted as the effects of peer achievement. A translation of the results shows that being surrounded by peers who score 1 point lower on average has the following effects: it lowers a black student's own score by 0.676 points in reading and 0.402 points in math; it lowers a Hispanic student's own score by 0.266 points in reading and 0.185 points in math; and it lowers a white student's own score by 0.168 points in reading and 0.092 points in math. The translation suggests that the effect of average peer achievement varies from small (0.092) to substantial (0.676) and that average peer achievement has its most substantial effects within racial groups.

In the 4th, 5th, and 6th grades only, Hispanic students perform worse in reading and math and white students perform worse in math when they are in classes with a larger share of Hispanic students. For instance, for every 10 percentage point rise in the share of their class that is Hispanic, Hispanic 5th graders' reading scores fall by 0.142 points and their math scores fall by 0.205 points. With the same change in the Hispanic share, white 5th graders' math scores fall by 0.061 points.

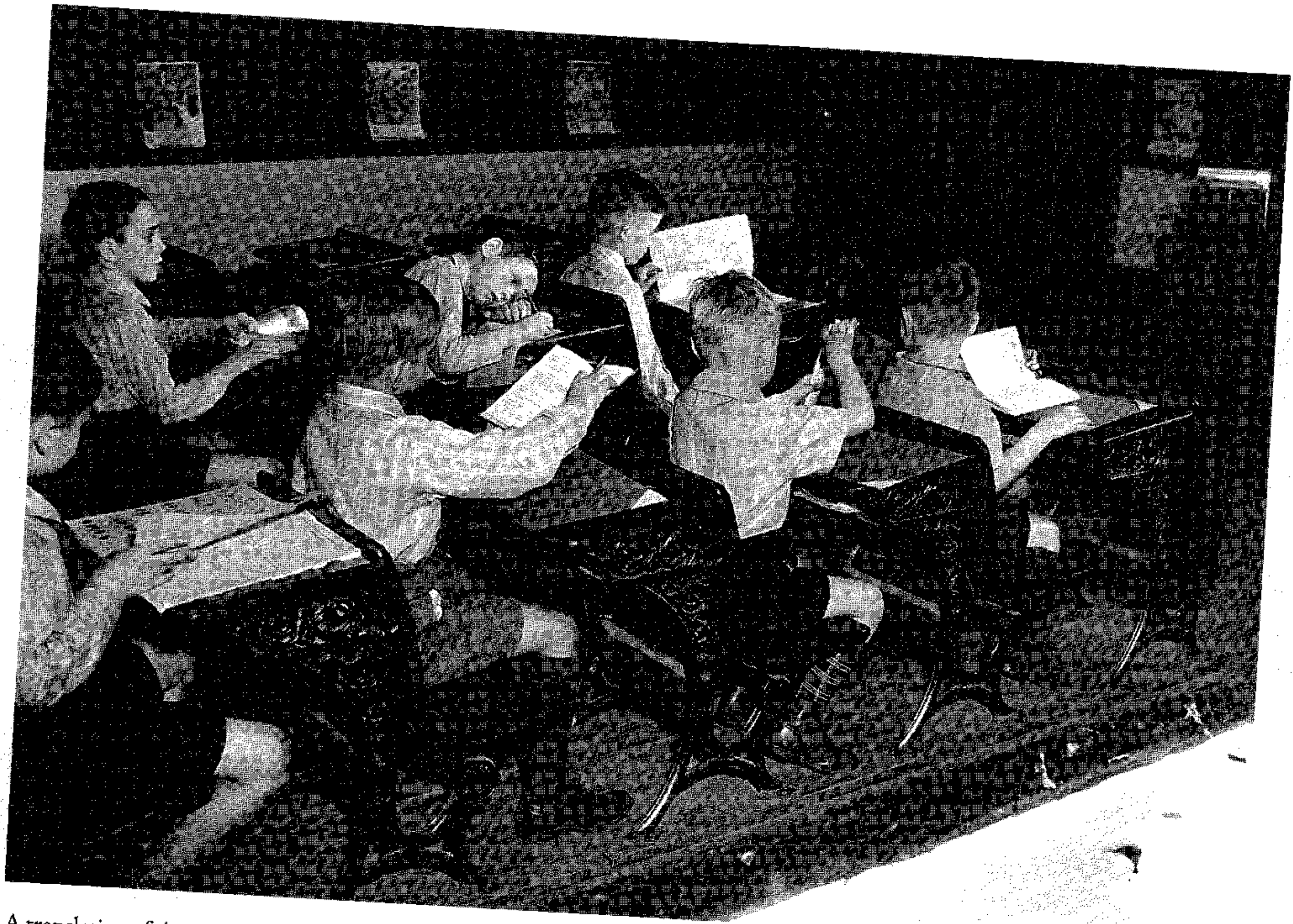
Girl Power (Figure 1)

As the share of girls in a classroom rises, the test scores of both boys and girls improve.



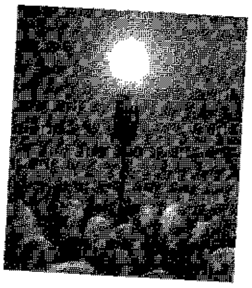
** Statistically significant at the .01 level
* Statistically significant at the .05 level

SOURCE: Author's calculations



A translation of the results finds that being surrounded by peers who score 1 point lower on average has the following effects: it lowers a Hispanic student's own score by 0.439 points in reading and 0.587 points in math, and it lowers a white student's own score by 0.176 points in math. Again, the results suggest that the effects of average peer achievement vary and are greatest for peers who are within the racial group that is generating the change in achievement.

The fact that peer effects appear to be stronger among members of the same race or ethnicity than across racial and ethnic groups suggests that the "baseline model" of estimating peer effects is inadequate.



There were a few results for Asian students that were statistically significant. Each of these results showed Asian students' having positive peer effects in math. For instance, with every 10 percentage point increase in the share of their class that is Asian, white 5th graders' math scores rise by 0.072 points and white 6th graders' math scores rise by 0.202 points. This comports with the interpretation that average peer achievement influences everyone's test scores, since Asians score higher than whites in math overall (the Asian-white score gap is positive and relatively large in math, 0.62 of a standard deviation in the 4th, 5th, and 6th grades).

The fact that peer effects appear to be stronger for members of the same race or ethnicity than across racial and ethnic groups suggests that the baseline model, in which the average achievement of one's peers has a linear effect on one's own achievement, is inadequate. Let's further explore the question of nonlinear peer effects by examining whether peer effects are different at various starting points: when the initial cohort is 0 to 33 percent black, 33 to 66 percent black, or 66 to 100 percent black.

Three patterns stand out. First, the negative peer effect of black students on black students' own scores is largest in cohorts that are between 33 and 66 percent black. The negative effect of black students on white students' own scores is largest in cohorts that are at least 33 percent black. I performed the same test with Hispanic students. The negative effect of Hispanic students on Hispanic students' own scores

PHOTOGRAPH BY HULTON/ARCHIVE

only appears in cohorts that are 0 to 33 percent Hispanic. In fact, Hispanic students have a statistically significant, *positive* effect on the achievement of Hispanic students in cohorts that are 66 to 100 percent Hispanic (see Figure 2).

There are a few possible interpretations of this last finding. First, having even more Hispanic peers in a cohort that is already mainly Hispanic may be helpful because each student who has difficulty speaking English is more likely to find a bilingual student to translate for him, to help him learn English, and so on. Second, an overwhelmingly Hispanic cohort may be helpful because it makes teachers sensitive to providing instruction that can be understood by students with limited English proficiency. Third, some schools, when faced with an unusually large Hispanic cohort, may segregate their Spanish-speaking students in a particular classroom because there are enough students to fill such a class. It is possible that such segregation generates higher achievement among Hispanic students (even if it is undesirable for other reasons). Regardless, this finding clearly shows that peer effects do not operate only in a linear fashion.

Results of Strategy 2

Remember that Strategy 2 uses cross-sectional data to study the impact on other groups when a particular gender or racial group experiences unusually high or low achievement. In the gender comparison, all of the results show that one gender's idiosyncratic achievement has a positive, highly statistically significant effect on the idiosyncratic achievement of its peers from the other gender group. In grades 3 through 6, being surrounded by peers who score one point higher in reading raises a student's own score by 0.3 to 0.4 points. Being surrounded by peers who score one point higher raises a 3rd grader's own math score by about 0.6 points, a 4th grader's own score by about 0.5 points, and a 5th or 6th grader's own score by about 0.4 points. In the racial and ethnic comparison, the results show that being surrounded by peers who score 1 point higher in reading raises a student's own reading score by 0.3 to 0.8 points. In general, the math results were similar to the results for reading.

In short, Strategy 2 generates unambiguous evidence about the existence of peer effects, but the range of estimates is somewhat wide: 0.10 to 0.55 points is a plausible summary of the range, given the various results and known biases.

Conclusion

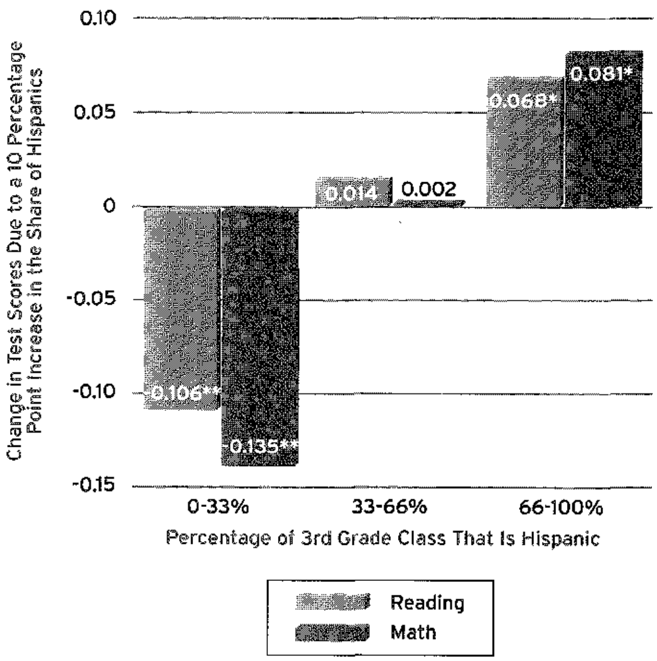
The peer effect estimates generated by the two strategies are reasonably similar. Strategy 1 found that being surrounded by peers who score 1 point higher raises a student's own score between 0.15 and 0.40 points. Strategy 2 tends to estimate an increase of between 0.10 and 0.55 points when a student is surrounded by peers who score 1 point higher. These estimates confirm that peers' ability levels affect achievement in ways that policymakers and researchers should not ignore.

Both strategies also showed that the baseline model of linear peer effects is inadequate. My results provide little evidence of general asymmetry, such as low achievers gaining more by being with high achievers than the amount high achievers lose by being with low achievers. However, I do show that peer achievement is not the sole channel for peer effects. The large, positive effect that a prevalence of girls has on boys' math scores cannot plausibly be explained solely by girls' effect on average peer achievement in math. Likewise, I found that a rising share of Hispanics has a positive effect on certain Hispanic students' scores, which could not be an effect of average peer achievement since raising the Hispanic share lowers average peer achievement. In addition, some results suggest that peer effects are stronger inside racial groups than between racial groups.

–Caroline M. Hoxby is a professor of economics at Harvard University and a visiting fellow at the Hoover Institution, Stanford University. The unabridged version of this article is available at www.educationnext.org

Critical Mass (Figure 2)

When Hispanics make up less than 1/3 of a class, an increase in the share of Hispanics leads to lower scores among Hispanics. But when a classroom is more than 2/3 Hispanic, an increase in the share of Hispanics actually improves their scores.



** Statistically significant at the .01 level
* Statistically significant at the .05 level

SOURCE: Author's calculations

Tracing the evolution of New American Schools, from feisty upstart to bulwark of the education establishment

unrequited promise

by JEFFREY MIREL

IN THE SUMMER OF 1991, in response to President George H.W. Bush's major education initiative, CEOs from a number of major corporations established the New American Schools Development Corporation (NASDC, later shortened to New American Schools, or NAS), a privately funded, nonprofit organization devoted to supporting the design and dissemination of "whole school reform" models. NASDC's founders envisioned a complete overhaul of American education stimulated by the spread of these innovative designs. As one put it, school reformers who hoped to receive NASDC grants had to "cast aside their old notions about schooling—to start with a clean sheet of paper, and be bold and creative in their thinking, and to give us ideas that address comprehensive, systemic change for *all* students for *whole* schools." To President Bush, NASDC represented a major step toward an "educational revolution" that would "seek nothing less than a new generation of schools."

The idea was to apply a research-and-development model to a sector that too often fell for romantic—and untested—notions of how schools and learning should be structured. NASDC's early leaders were determined to apply a no-nonsense business approach to their efforts, to create an organization that was as lean and agile as the corporations they led. New American Schools would be less bureaucratic and more aggressive in responding to new ideas than the typical government agency or major foundation. Moreover, NASDC would be a sort of venture capitalist for education, constantly evaluating its investments and continuing to fund only those designs that proved their effectiveness.

ILLUSTRATION BY ANITA KUNZ

Initially, the plan for NASDC was to complement the creation of 535 federally funded "break the mold" schools by 1996. Each of the 435 congressional districts in the country would be home to one of these schools; two more schools per state would be funded as well. Each school would receive a one-time federal grant of $1 million in start-up funds. The privately

funded NASDC would support R&D teams tasked with designing innovative reform models to be implemented in these schools.

NASDC's original plan didn't include supporting the implementation of these designs. But as the political environment soured early on, it became increasingly clear that NASDC wouldn't have 535 federally funded laboratories in which to test its designs. In light of this, NASDC had to become more than just an R&D center; it would have to support not only the R&D efforts of the design teams, but also the dissemination and implementation of their ideas. Moreover, if it was to keep its promise of acting as a kind of venture capitalist for education reform, at some point these design teams would need to become self-sustaining, taking fees from school districts in return for their expertise.

Over a decade, this mandate would transform New American Schools from a privately funded R&D center that would support only effective whole-school reform models into a savvy service provider for research teams that needed to market and fund their designs. Their ability to sell those designs, in turn, was heavily dependent on the ability of New American Schools to lobby Congress for more federal funding of whole-school reform. In some ways, these changes strengthened NASDC's claim of being a bold new venture in education reform, since few privately funded initiatives had ever committed to so ambitious a program. But along the way New American Schools lost sight of its mission and became more of a typical Washington insider than an education revolutionary.

Progressivism Rides Again

NASDC's initial funding competition sought innovative ideas for creating schools in which all students would reach "world class standards in English, mathematics, science, history, and geography" and be "prepared for responsible citizenship, further learning, and productive employment." The "request for proposals," or RFP, process also required that the designs be replicable in other communities: "*This is not a request to establish 'model' schools.* The designs must be adaptable so that they can be used by many communities to create their own schools. The important thing is that long after NASDC has disappeared from the scene, its legacy for new designs will remain." NASDC funds would support the start-up of these break-the-mold schools, but once they were running it was expected that they would operate on the same budgets as conventional schools.

Once NASDC chose the 11 winning proposals (see sidebar below), it established a three-phase process for refining, imple-

The Original New American Schools Designs

The competition for New American Schools funding yielded 11 winning designs that were notable mainly for their use of progressive rhetoric in describing their approaches.

- **Audrey Cohen.** Organized as a series of "complex and meaningful . . . purposes" in which students apply content knowledge to solving "real world" problems. The curriculum is interdisciplinary and project-driven, with the goal of having schoolwork contribute to "improving the world outside of school."

- **Authentic Teaching, Learning, and Assessment for All Students (ATLAS).** Drawing on the work of several prominent educators, such as Theodore Sizer, James Comer, and Howard Gardner, ATLAS engages students in "active inquiry" as part of an integrated curriculum built around "essential questions." The goal is to create "communities of learners."

- **Bensenville New American School.** Bensenville planned to use "the entire community as a school campus," with pupils actively involved in real problems in diverse settings (for example, students could be assigned to a local business to see how "learning has a tangible and immediate purpose"). Students were grouped according to their progress in moving through an interdisciplinary curriculum rather than by age or traditional grades.

- **Community Learning Centers of Minnesota.** Under the aegis of Minnesota's new charter-school law, this design sought to empower teachers, parents, and community members to create schools that are "learner centered" and "meaningful." The key to finding meaning was involving students in "real life tasks" such as "using statistics to conduct and analyze a community survey of attitudes about taxes."

- **Co-NECT Design for School Change.** Constructed around an interdisciplinary curriculum that combined ongoing student-initiated projects with teacher-initiated seminars. Supported by a number of prominent high-tech companies (Apple, Lotus, NYNEX), this design has a strong emphasis on students' using technology for projects and for online interchanges among members of the school community.

menting, and marketing these designs. Phase one provided one year of funding, during which the teams fine-tuned their original proposals "into workable designs for school transformation." At the end of phase one, NASDC dropped two of the designs, the Bensenville New American Schools and the Odyssey Project. Phase two provided an additional two years of NASDC funding for the design teams to implement (and further refine) their ideas and approaches. At the end of phase two, NASDC determined that two more designs, Community Learning Centers and Los Angeles Learning Centers, would not receive further support (although the Los Angeles group eventually rejoined New American Schools with a new name, Urban Learning Centers, and a more focused mission). Finally, phase three envisioned a two-year scaling-up process in which the remaining designs would be put into practice in a "large number of schools across the country." NASDC stuck to its timetable for phases one and two, but the scale-up phase, which began in mid-1995, has taken longer and, in essence, continues today.

The popular and media response to the NASDC initiative was generally upbeat and positive, but a small, vocal group of critics repeatedly questioned NASDC and the designs. Initially the negative assessments came primarily from liberal critics of the Bush administration. Typical was a 1992 article in the *Nation*, in which the authors declared: "Most of the educational R&D teams endorsed by the corporation comprise an incestuous circle of right-wing ideologues and privatization advocates, teacher-hating technocrats and recession-rocked military contractors, their funding made palatable to the press by token support for established and respected liberal school-reform advocates."

Others argued that NASDC had misplaced priorities and ignored such pressing issues as equalizing funding for poor districts and addressing the needs of minorities. Former U.S. commissioner of education Harold Howe II, for example, derided NASDC's business-oriented approach, declaring, "You can bet that, when these parties get together, they will spend little time in making a diverse society work."

Historians David Tyack and Larry Cuban wondered whether the New American Schools initiative was just the latest example of "ambitious reformers" promising "to create sleek, efficient school machines 'light years' ahead of the fusty schools of their times," yet ultimately being overcome by the education system's unreceptive environment. They argued that such efforts to reinvent schools have in practice "often resembled shooting stars that spurted across the pedagogical heavens, leaving a meteoric trail in the media but burning up and disappearing in the everyday atmosphere of the schools."

These criticisms were provocative, but ultimately proved less than prophetic. Contrary to Howe's prediction, the New Amer-

- **Expeditionary Learning.** Based on the principles of the Outward Bound program, this design blends intellectual, emotional, and character education into a single curriculum. The course of study is based on the international baccalaureate, but its strong emphasis on content was "intertwined" with engaging learning expeditions, on such topics as "Understanding the Bubonic Plague."

- **Los Angeles Learning Centers.** Structured as a "moving diamond" that linked each child in the program with three different mentors–an older student, a parent or community member, and a teacher. This foursome stayed together for several years to provide academic and emotional support for the child.

- **Modern Red Schoolhouse.** Seeks to fuse the best of old-fashioned one-room schools (for example, multi-age homerooms that keep students and their teacher together for several years) with the use of information technology. It focuses on student mastery of a strong, academic curriculum (influenced by the ideas of such educators as E. D. Hirsch Jr. and William J. Bennett).

- **National Alliance for Restructuring Education.** Concentrates on both systemic and school-level reforms. The design coordinates district-wide education, health, and human services in ways that aim to improve student learning. Schools are run according to the principles of Total Quality Management.

- **The Odyssey Project.** This community-based design developed in Gaston County, North Carolina, dispensed with traditional grades and replaced them with broad groupings ("Alpha" for ages zero to three, "Beta" for ages three to six, and so on). Students progressed at their own pace through the interdisciplinary curriculum by meeting specific "performance outcomes." Older students participated in weekly seminars that focused on "national and world citizenship" and multicultural issues.

- **Roots and Wings.** Building on Robert Slavin's Success for All program, this design focuses on having students develop strong basic skills such as reading and writing (the roots) as well as content knowledge and higher-order thinking skills (the wings). A central feature is the WorldLab, which involves students in real-world activities, projects, and simulations, such as representing the 13 colonies in debating and drafting the U.S. Constitution.

ican Schools designs have focused overwhelmingly on improving education for urban and minority students. And NASDC and its design teams, rather than "disappearing in the everyday atmosphere of the schools," have survived for a decade, mainly due to their ability to adapt to the changing political and pedagogical environment.

Ironically enough, the attacks from conservatives have proven more prescient. They maintained that the RFP process was skewed to ensure the choice of safe rather than break-the-mold designs. The competition was more like those run by government agencies or major foundations than an agile start-up venture. This gave an advantage to experienced grant writers, and most of the winning proposals came from established education reformers and their organizations.

In its current incarnation, New American Schools is almost indistinguishable from the groups that have dominated school reform for decades. It has become part of the very education establishment its leaders once sought to circumvent.

The competition rounded up not only many of the usual educational suspects but the usual educational ideas, too. Lynn Olson of *Education Week* pointed out that one "striking feature of the award-winning designs is how many ideas they have in common." A more striking feature was how many of those ideas were deeply influenced by the questionable theories of the progressive education movement of the early 20th century, the original break-the-mold movement in American education. Indeed, most of the winning proposals read like echoes of progressive manifestos of the 1920s. Six of the eleven designs, for example, expressly declared that they were "learner-centered." Eight sought to change the relationship between teachers and students by transforming the teacher into a "coach," "facilitator," or "guide." All promised to meet "world class" curriculum standards, but only two (Expeditionary Learning and Modern Red Schoolhouse) focused their programs on students' explicitly mastering academic disciplines. Nearly all the designs promoted interdisciplinary curricula to avoid what they saw as a key problem of traditional schools: teaching knowledge and skills in isolation.

Related to this interdisciplinary focus was the widely shared commitment to use the progressive-inspired "project method" to engage the interest of students. Eight NASDC designs proclaimed that project-based education would make learning exciting and relevant to students in ways that traditional textbook-centered education never could. Eight proposals envisioned using the community as a classroom to break down barriers between school and society, an idea championed by John Dewey, the great patriarch of progressive education. Like Dewey and other progressives, most of the proposals also argued that their design would not lead to the "piling up of facts," but would produce critical thinkers.

The problem, as critics such as Robert Maynard Hutchins, E. D. Hirsch Jr., and Diane Ravitch have repeatedly shown, is that apart from a small number of schools serving highly motivated, well-to-do students, progressive education has rarely fulfilled its pedagogical promise. These critics have argued that progressive ideas often contribute to the evisceration of academic curricula and narrow the scope of student learning to entertaining but largely vapid topics. Especially relevant to the NASDC designs was the late Jeanne Chall's finding, after surveying a host of research studies, that, on average, schools guided by progressive ideas have been less successful in raising academic achievement, especially among disadvantaged children.

It was, in fact, controversies over progressive ideas that were at least partly responsible for the Bensenville and Odyssey Project designs' collapse in the face of angry community challenges. Both designs were grassroots efforts developed by local communities; neither was guided by or associated with the work of a national education leader. Still, among the most contentious elements of both designs was their progressive reorganization of the schools. A sociological study of the North Carolina–based Odyssey Project found that parents were bothered by the lack of course grades and college prep courses in the Odyssey curriculum. They saw Odyssey as an attempt to shift the curriculum to a more vocation-oriented track that would ensure a "compliant, trained workforce" for local industry.

These developments should have sent a warning to NASDC leaders about the politically charged and therefore tenuous nature of such broad efforts at education reform. These debacles showed that the success of NASDC-style reform—no matter how good the design nor how much money available—would depend ultimately on keeping all the community's stakeholders committed to change over the long haul.

Regardless, phase one had gone generally well for the other nine design teams, and NASDC leaders plowed ahead undaunted. As businessmen, they recognized that any new venture has some failures and losses. More troubling for them at the time were two emerging problems—NASDC's precarious financial situation and concerns about the organization's survival following the defeat of George Bush by Bill Clinton in 1992.

Scale-Up

In its first two years of operation, NASDC raised only about $50 million of a planned $200 million. Concerns about fund-raising deepened with fears that NASDC was too closely tied to the Bush administration to win the new president's support.

In May 1993, however, both President Clinton and Secretary of Education Richard Riley stated that the NASDC's reform efforts dovetailed nicely with the new administration's education agenda. In fact, this was no real stretch for Clinton. He had represented the nation's governors at the 1989 meeting at which Bush and the governors had agreed on this agenda, and his administration embraced many of the same goals.

Yet Clinton's support had little immediate effect on fund-raising, and NASDC watchers began asking whether the organization could remain in business much longer. In response, the NASDC board elected David Kearns CEO. Kearns, who had previously served as deputy secretary of education in the Bush administration and as CEO of Xerox, was widely respected in political, education, and business circles. The board hoped he would find ways to shore up the organization's precarious financial situation. In January 1994, these hopes were borne out when philanthropist Walter Annenberg donated $50 million to NASDC—a life-saving development for the organization. Without these funds it was unlikely that NASDC could have made the crucial transition from serving a modest number of schools to its broader scale-up to the state and district level.

Further enhancing NASDC's national profile was the fact that its key scale-up idea—moving from reforming individual schools to transforming entire systems—was quite similar to the idea of systemic reform that was one of the cornerstones of the Clinton administration's education policy. NASDC was now freed from the allegation that it was little more than a partisan GOP initiative. In the process, however, NASDC also appeared to be distancing itself from its "revolutionary" free-market approach to education reform and was increasingly looking like a respectable if rather corporate member of the education establishment.

With its status bolstered, NASDC turned to its biggest challenge: scale-up. Between 1992 and 1995, nine NASDC-supported design teams had implemented their plans in 148 schools. Phase three called for a dramatic increase, with projections reaching into the thousands. At the time, however, none of the designs had demonstrated that it would lead to dramatic improvements in student achievement (although Robert Slavin's Success for All program, the cornerstone of the Roots and Wings initiative, had shown promising results).

Nevertheless, NASDC moved forward with its scale-up plan. Agreements were ultimately reached with the states of Kentucky and Maryland and nine large urban districts: Cincinnati, Broward County (Fort Lauderdale), Dade County (Miami), Memphis, Philadelphia, Pittsburgh, San Antonio, San Diego, and

Unproven Designs

Only 3 of 24 popular models of whole-school reform have strong evidence that they improve student achievement. None of the 3 is a New American Schools design.

Models of Whole-School Reform	Evidence of Positive Effect on Student Achievement	Year Introduced	Number of Schools Adopting the Model*
Accelerated Schools (K-8)	Marginal	1986	1,000
America's Choice (K-12)	No Research	1998	300
ATLAS Communities (PreK-12)	No Research	1992	63
Audrey Cohen College (K-12)	No Research	1970	16
Basic Schools Network (K-12)	No Research	1992	150
Coalition of Essential Schools (K-12)	Mixed, Weak	1984	1,000
Community for Learning (K-12)	Promising	1990	92
Co-NECT (K-12)	No Research	1992	75
Core Knowledge (K-8)	Promising	1990	750
Different Ways of Knowing (K-7)	Promising	1989	412
Direct Instruction (K-6)	Strong	Late 1960s	150
Expeditionary Learning Outward Bound (K-12)	Promising	1992	65
The Foxfire Fund (K-12)	No Research	1966	N/A
High Schools That Work (9-12)	Strong	1987	860
High/Scope (K-3)	Marginal	1967	27
League of Professional Schools (K-12)	Marginal	1989	158
Modern Red Schoolhouse (K-12)	No Research	1993	50
Onward to Excellence (K-12)	Marginal	1981	1,000
Paideia (K-12)	Mixed, Weak	1982	80
Roots and Wings (Pre-K)	Marginal	1993	200
School Development Program (K-12)	Promising	1968	700
Success for All (PreK-6)	Strong	1987	1,130
Talent Development High School (9-12)	Marginal	1994	10
Urban Learning Centers (PreK-12)	No Research	1993	13

**As of October 30, 1998*

SOURCE: "An Educator's Guide to Schoolwide Reform," American Institutes for Research, 1999, under contract to AASA, AFT, NAESP, NASSP, NEA

Seattle (and a group of contiguous school systems). By mid-1997, two years into phase three, the number of participating schools had climbed to 553.

Memphis would become NASDC's signature district. In 1997 its student population was more than 80 percent African-American, and more than 60 percent of all students qualified for free or reduced-price lunch. Only 40 percent of its 9th graders were passing the state's minimum-competency test. However, Superintendent Naomi "Gerry" House was determined to use whole-school reform to try to solve Memphis's education ills. By 1997, Memphis had almost reached the NASDC goal for jurisdictions participating in the scale-up of having 30 percent of its schools adopting some form of whole-school design. Encouraging news about student achievement soon followed. A study comparing 25 NASDC school sites in 1995–96 with a control group of non-NASDC schools found "significantly greater gains" in the redesigned schools.

Several early debacles should have sent a warning to New American Schools' leaders about the politically charged and therefore tenuous nature of such broad efforts at education reform.

Meanwhile, less encouraging news was emanating from the RAND Corporation's studies of NASDC's phases one and two. RAND, with which NASDC had been working from the beginning, found that one of the main problems in the early stages was the grating between these innovative designs and the superstructure of state and district policies. This spelled trouble for the scale-up phase, when entire systems would be expected to convert to these new designs while keeping such major influences as union-negotiated contracts and state policies largely intact. For instance, in a RAND survey of principals in NASDC-supported schools, a large majority reported that "standardized, multiple-choice tests are misaligned with the classroom practices of reforming schools." This is a common criticism among all educators, but here the NASDC principals were voicing a more pointed concern, namely that there was a mismatch between the student-centered approaches of the design teams and the more traditional evaluation practices used by states and districts.

Moreover, a 1995 survey of parents in several NASDC jurisdictions found that they were deeply divided about reform, with over a third of the respondents wanting "the best of new and innovative methods"; just under a third wanting "tried and true" basics; and the remaining third wanting "a combination of old and new methods." Getting such communities to commit to long-range plans for whole-school transformations would be a formidable challenge. This indicated a process of change that was more incremental and political than the NASDC teams had expected, as they would soon learn in Memphis. The early phase indicated that NASDC's original perspective was somewhat simplistic, even naive. NASDC could not simply apply the business metaphor—"develop and test a new product and then go sell it"—to educational change.

As a result, by 1995, NASDC leaders rarely spoke about creating break-the-mold schools; the words "education revolution" had disappeared entirely. The realities of education politics were forcing NASDC to adapt and adjust its approach and expectations.

Congress to the Rescue

By mid-1997, NASDC, now named New American Schools, or NAS, was well into becoming more of a clearinghouse than an initiator of reform and was changing the design teams into market-oriented purveyors of education improvement. NAS increasingly focused on helping the teams become self-sufficient organizations capable of marketing and supporting their products. For some teams, this change was relatively painless. Success for All, the parent organization of Roots and Wings, was already a fee-for-service institution with substantial experience in this area. But others, such as the Modern Red Schoolhouse, had virtually no business or marketing experience. Thus, during phase three, NAS began sending financial and marketing experts to the design teams to help them develop business plans and marketing strategies. At the same time, NAS began addressing the problem of creating markets for the designs.

The most immediate problem for NAS and the teams was finding funds that districts could use to pay for their services, since design-team fees for the first year of implementation alone could range from $50,000 to more than $500,000 per school. This effort was especially crucial given that so many of the clients targeted by NAS were low-performing, financially strapped urban systems.

Help came in the form of the Comprehensive School Reform Demonstration Program, better known as the Obey-Porter Amendment. Passed by Congress in 1997, it provided about $150 million for competitive grants to aid schools in adopting "proven" whole-school reform models. Supporters of the amendment routinely cited the NAS designs as representative of the kinds of programs they hoped school districts would adopt. In fact, NAS marketing efforts and Obey-Porter funds have brought a five-fold increase since 1997 in the number of schools using NAS designs, to some 3,000 schools in 2001. Not surprisingly, NAS had lobbied Congress on behalf of the amendment.

One could view Obey-Porter as helping NAS fulfill its original role, namely acting as a private R&D agency whose designs would be implemented in schools that received federal start-up funds. Yet if Obey-Porter aided NAS in fulfilling an important goal, it also indicated that NAS was changing from an agency committed to operating outside the constraints of Washington into one that depended on Washington for the success of its venture.

However, the most troubling aspect of Obey-Porter and, by implication, of the new role played by NAS design teams in education reform, was not the dependency on federal funding but the claims of "proven effectiveness" that the law explicitly made on behalf of the designs. With the possible exception of Roots and Wings, at the time that Obey-Porter was enacted the NAS designs still couldn't provide substantial evidence of effectiveness.

By 1995, the leaders of New American Schools rarely spoke about creating break-the-mold schools; the words "education revolution" had disappeared entirely.

Disappointing Results

Indeed, NAS officials' reaction to the 1999 American Institutes of Research (AIR) study *An Educator's Guide to Schoolwide Reform* would reveal an organization that had not come to terms with the importance of evaluating student achievement in the reform process. Of 24 whole-school reform designs, AIR found that only three had "strong" positive effects on student achievement: Direct Instruction, High Schools That Work, and Success for All, none of which were NAS designs. Of the eight NAS programs, AIR rated Expeditionary Learning "promising" and Roots and Wings (the NAS version of Success for All) "marginal" in boosting achievement. All the others received no rating because there weren't enough data.

New American Schools president John Anderson's reaction, in a 1999 column for *Education Week*, was to maintain that the AIR review was flawed in its approach and conclusions. "Not surprisingly," he argued, "AIR took a traditional approach, gathering the relatively small body of published research on the programs as its evidence of effectiveness. But this sort of reliance on limited and often time-consuming research as a way of assessing comprehensive designs may be outmoded." Anderson held that "traditional" evaluation took too long and ignored data on progress that schools were gathering on an almost daily basis. In addition, he wrote, "Even the most careful research on a handful of campuses and a handful of matched schools could not give us a picture of what happens when models are used in hundreds of schools." Anderson seemed troubled by the attention AIR gave to student achievement in making its assessments. "Beyond academic results," he asserted, "design teams should be judged on other practices involving their relationship with schools and school districts. Practices involving implementation, communication, and data collection are all vital to viable campus change."

Was Anderson implying that the NAS designs were so novel that standard approaches to research could not capture what they were doing? Was he questioning all traditional external reviews of the NAS programs? If so, these concerns didn't stop him from citing data from a traditionally designed study on student achievement in Memphis as evidence that the designs in that city were working.

Meanwhile, the data finally produced by RAND on the performance of NAS schools in scale-up jurisdictions relative to other schools in the same district showed mixed results. About half of the NAS schools in Cincinnati, Memphis, Miami, and Kentucky (the cities where RAND studied math performance) outperformed the district average in mathematics, while 50 percent or more of the schools in Cincinnati, Philadelphia, San Antonio, and Washington (the cities where RAND studied reading performance) improved in reading. In other words, about half of the NAS schools in the cities studied by RAND only met the district average or underperformed relative to other district schools. RAND identified Roots and Wings as the most consistent design, with 10 of 21 schools making progress in both reading and mathematics relative to the district. Modern Red Schoolhouse did even better, with 7 of 11 of its schools improving in math and 8 in reading. Other designs showed considerable variation. Of the Audrey Cohen schools, for example, 5 of 8 schools improved in math, but only 2 in reading. Similarly, 9 of 24 ATLAS schools and only 4 of 16 Expeditionary Learning schools improved in math.

The inescapable finding from RAND was that, despite millions of dollars and enormous effort, over the course of a decade what NAS had brought about was at best incremental change and modest achievement gains. RAND's mixed assessment was: "Dramatic achievement gains were not made, although some designs fared better than others."

Another blow to NAS's performance record was dealt in June 2001, when the Memphis school district announced that it was abandoning the district's six-year experiment in whole-school reform. In the spring of 2000, when Johnnie Watson took over as superintendent from Gerry House, the great promoter of whole-school reform, he found a deeply troubled district. In

the late 1990s, House had mandated that *all* of Memphis's more than 160 schools adopt a reform model, a policy that angered and alienated many teachers. Worse, an internal district study found that after six years of reform and the expenditure of some $12 million, Memphis students had exhibited virtually no gains on state tests. Moreover, the study revealed that only 3 of 18 reform models implemented in the city had boosted student achievement; none of the three was a NAS design. Indeed, one NAS design, ATLAS, had "the most negative impact" on student achievement of any of the reform models. The report found that even Success for All, which was being used in almost 25 percent of the Memphis schools, performed poorly. There have been questions about the quality of the study, but this merely underscores one of NAS's fundamental problems: the lack of independent, reliable research on the effectiveness of its designs, a deficiency that allows decisions like Memphis's to be made more on the basis of politics than sound evidence about what works.

These events should signal to NAS that the future of whole-school reform rests as much on the fluid and unpredictable nature of educational politics, changes in leadership, and the sufferance of parents and teachers as it does on successful marketing, buy-in from key administrators, and substantial federal funds. Moreover, the Memphis experience highlights the importance of the bottom line—student achievement—for all such reform efforts.

How will Memphis's rejection of whole-school reform affect NAS? While it is certainly too soon to offer a definitive answer, it seems likely that NAS will weather this storm, although its momentum may be somewhat slowed. The flow of Obey-Porter and, increasingly, Title I funds to support whole-school reform will sustain NAS and its design teams for many years to come.

The most troubling aspect of Obey-Porter was not the dependency of New American Schools on federal funding but the claims of "proven effectiveness" that the law explicitly made on behalf of the NAS designs.

Lost Chances

At its inception New American Schools held the promise of being an extremely exciting research-and-development initiative in education. It would sponsor the creation of innovative designs, pilot test them in a select group of schools, and decide whether they were effective enough to warrant wide dissemination. Its evolution into a financier, lobbyist, and marketing shop for a variety of as-yet-unproven whole-school designs must be considered at best a modest contribution to the cause of education improvement. In its current incarnation, NAS is almost indistinguishable from the foundations and government agencies that have dominated school reform efforts for decades. It has become part of the very education establishment its leaders once sought to circumvent.

More important than the changed status of NAS is the increasing prominence of NAS-style whole-school reform as one of the most popular strategies for education change. NAS has played a key role in that development. Yet its own experience in Memphis raises questions about whether the commitment to whole-school reform is the best strategy for improving educational outcomes in troubled urban schools. With so many unmanageable and unpredictable variables, whole-school reform will always be tenuous, because it attempts to simply glide by many of the institutional constraints—the variety of programs operating within a school, all with different goals; the requirements of the administrators' and teacher unions' contracts—that make large-scale change so tortuous. University of Michigan historian Maris Vinovskis and other scholars have suggested that less ambitious and less costly efforts such as more learning time during the summer might yield gains as great as or greater than comprehensive school reform.

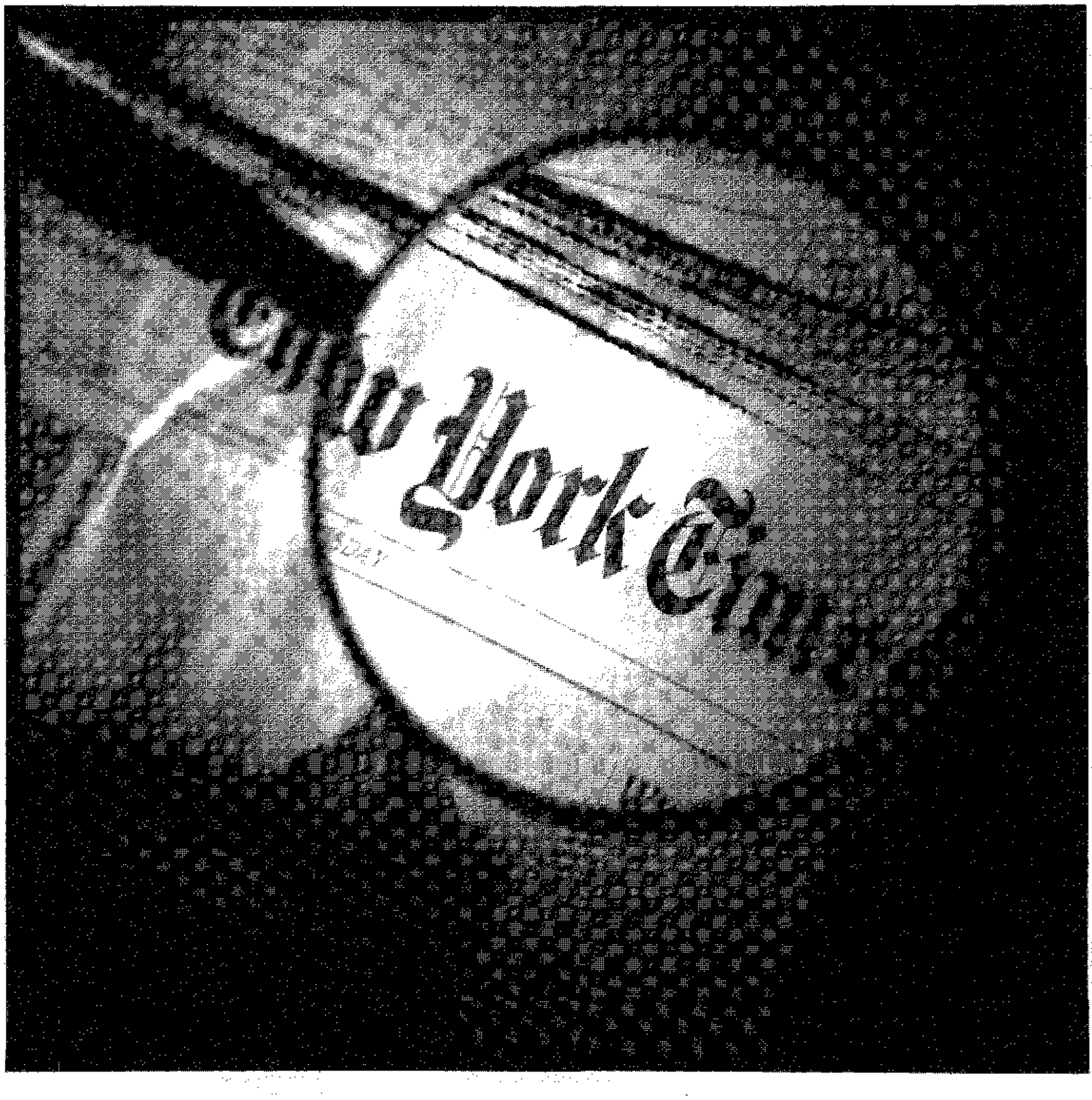
One of Johnnie Watson's first moves after ending Memphis's experiment with whole-school reform was to order a new district-wide curriculum. A primary reason was the problem the district had experienced in adopting so many different designs. One teacher noted: "Since Memphis has such a mobile student population students fell even further behind because nothing was the same from school to school." High rates of student mobility are an ever-present feature of all urban school districts. Yet the NAS design teams did not appear to have taken that vital fact of urban life and education into consideration when they developed their programs. The seeds for significant gains may lie in such small insights about student mobility rates and district-wide curricula.

–Jeffrey Mirel is a professor of educational studies and history at the University of Michigan and the author of The Rise and Fall of an Urban School System: Detroit, 1907–81 (University of Michigan Press, 1993). *The unabridged version of this article is available at www.educationnext.org.*

check the facts
by DAVID W. MURRAY

Waiting for Utopia

The New York Times education columnist believes in education reform. He just doesn't think it has much to do with schools



PHOTOGRAPHY BY WEBB CHAPPELL

IT'S EASY TO TELL when someone is in the grip of a Big Idea That Explains Everything. Tunnel vision sets in; every analysis, whatever the topic, becomes an occasion for the grand theory to appear. Evidence is read and supplied selectively, in such a way that the theory remains unscathed. Skepticism is deployed selectively as well. Findings that comport with the Big Idea are held to a relaxed standard, while the work of critics is subjected to withering scrutiny.

Richard Rothstein, author of the *New York Times*'s widely read "Lessons" column, a weekly commentary on education issues, frequently exhibits these symptoms. His Big Idea is that economic forces, especially inequality and poverty, largely determine the outcome of American social projects—including attempts at education reform. The effect of this obsession is two-fold. First, his writings display a factual carelessness, suggesting that details hardly matter if one possesses a higher truth. Second, nearly every engagement with issues of schooling, testing, standards, and teaching becomes an occasion to reassert the primacy of the economic factor.

Rothstein believes that most contemporary criticism of the public schools is misplaced. The main problems lie not with the schools, he claims, but with the injustices associated with the American economic system. It isn't a lack of competition in the public school system, antiquated hiring and compensation systems, or a dearth of solid research on educational methods that depresses student achievement; it's an economic system that allows for large differences in income and wealth. As a result, to Rothstein's mind, education reform won't be effective without far broader social reforms.

The Sky Isn't Falling

Rothstein is committed to the view that no crisis exists in American education, that all the critics are merely Cassandras trying to scare the public into accepting their pet reforms. His statistical gymnastics are used most often in the service of defending current school practices against most reform proposals, usually by denying the need for reform in the first place. Thus Rothstein's explanation for falling SAT scores is not a decline in education quality, but an increase in the number and the diversity of test takers. Says Rothstein, in an August 2000 column: "Interpreting the SAT is more complex than it seems. SAT trends would reflect school quality changes only if every 18-year-old took the test. Not all do. Average scores are affected by who takes the SAT. If only the brightest seniors take it, averages are higher. If more lower-ranked seniors aspire to college and take the test, this could indicate better performance by schools, but still depress the average."

The claim that slipping scores result from a changed demographic (and hence could even be good news) has surfaced repeatedly in the writings of education commentators such as Gerald Bracey, but it is demonstrably false. *Washington Post* economics columnist Robert Samuelson summarized the matter in a 1994 column by noting: "The change in the student population *preceded* the drop in test scores. Between 1951 and 1963, the number of test takers went from 81,000 to nearly 1 million; test scores rose slightly." Moreover, the percentage of test takers remained relatively constant between 1972 and 1984 (see Figure 1). There were still a million test takers in 1985, the first year in which test scores showed a small uptick after 19 years of decline. Scores have been flat or slightly improved since then, with math scores returning to their levels of 30 years ago, but failing to reach their mid-1960s apex.

The main problems lie not with the schools, Rothstein claims, but with the injustices of the American economic system.

Changes in the composition of the test-taking pool don't explain the decline in test scores either. Studies by the Educational Testing Service and others have showed, in the words of Robert Samuelson, that "the main declines occurred among whites and could not be explained by changes in student's gender, economic class, or parental education." This analysis was seconded by Harvard sociologist Christopher Jencks, who pointed out that the SAT scores of advantaged white males have also exhibited a steep decline.

Yet Rothstein, exactly one year later, parroted his earlier claims. His reaction to the release of the 2001 scores, which showed no improvement over the previous year and hence were termed "stable" by the College Board, was to write, "Stable in this case does not mean unimproved. Hidden in the data is more hopeful news than most people would expect. These tests are voluntary. If only high achievers take them, average scores mean one thing. But if a broader range of students takes them, the results must be interpreted differently. The number taking the tests has in fact grown a lot.... It is remarkable that averages gained at all while the test-taking base was expanding."

Again, in an October 2001 column, the familiar refrain reappeared in a rebuttal of the 1983 *Nation at Risk* report. The authors, we are told, "misunderstood the decline in test scores. College Board results had dropped, but that was due to the growth in college-going ambition. The SAT was no longer taken only by top students, and so average scores of test takers naturally fell."

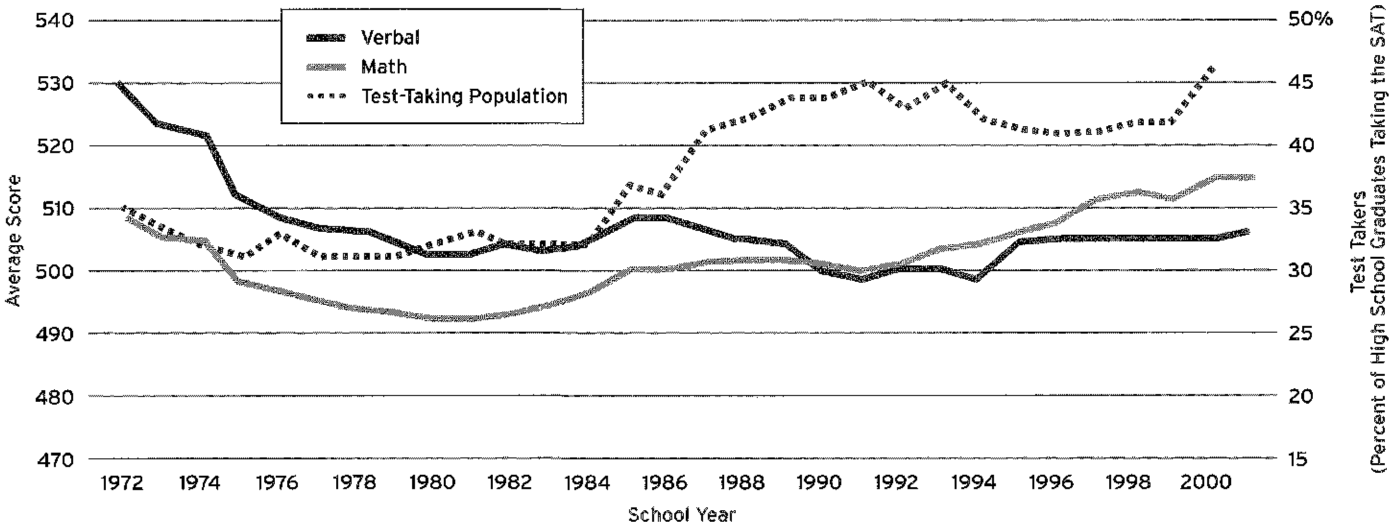
In short, in three successive bombing runs, Rothstein failed even to acknowledge, much less refute, any evidence that would undermine his assertions. This suggests an opinion resistant to the complexities of the issue.

Having dismissed claims that the achievement of American students has declined over time, Rothstein turns to the other major source of worry: the woeful performance of American students on international comparisons of educational achievement. Take a May 2001 column, in which Rothstein manages to convert lead into silver. He starts by facing the facts. He concedes that on the Third International Mathematics and Science Study, "Our 8th graders scored below their peers in almost every other industrial nation that took part. Students in Japan and Korea ranked near the top. In math, average American 8th graders would have scored below the 25th percentile in Japan or Korea." But don't worry, because the kids certainly aren't: "The study also asked students if they liked math and science.... here Japanese and Koreans scored at the very bottom.... In the United States, 35 percent felt positively about math and 32 percent about science, more than in almost every other industrial nation." A Japanese scholar is also invoked to assure us that his countrymen do "not attach great importance to students' rankings because the exams measure skills valued by the old education system, not the new." In fact, Rothstein concludes, the dour Japanese want to emulate our schools because of our "zest for living."

In a July 2001 column, Rothstein tells his readers not to fret over data from the National Assessment of

Misconstrued Trends (Figure 1)

SAT scores actually fell while the test-taking population hovered at around 33 percent through 1984. Since then, the increase in scores has accompanied a rise in the percentage of students taking the test, belying the claims of Rothstein.



SOURCE: Digest of Education Statistics, 2000, National Center for Education Statistics; The College Board

Educational Progress showing that two-thirds of American 4th graders can't read above a basic level, because "on an international survey of reading ability, American 4th graders scored higher than pupils everywhere except Finland." In other words, international comparisons are apparently valid when they corroborate Rothstein's fundamental beliefs, but easily dismissed when they reflect poorly on the American education system.

Poverty's Pull

For Rothstein, even if there were a crisis in American education, it certainly wouldn't be the schools' fault. Any problems that do exist are the result of social inequality. Concerned about low SAT scores? Rothstein insists that "the best predictor of test scores has always been students' social class." Thus low scores are to be expected from disadvantaged youngsters.

Rothstein returns to this theme time and again. For instance, Rothstein's December 2001 column highlighted a study by Eric Dearing of the Harvard Graduate School of Education. Dearing examined a ratio termed the "income-to-needs" of families below the poverty line and showed that when income increased (roughly $4,500 per year over three years), very young children performed better on tasks where they were asked to identify colors, shapes, and letters (skills considered important in school readiness).

Contrary to Rothstein's claims, changes in the composition of the test-taking pool don't explain the decline in SAT scores.

This was a only a small-scale study, but that didn't stop Rothstein from drawing the far-fetched conclusion that "educators have a stake in promoting a federal income policy that focuses on immediate income support for the unemployed, because this in itself could make instruction more effective." There's nothing wrong with arguing that the unemployed should receive more generous benefits, but this has hardly anything to do with education; it's certainly not a conclusion based on solid evidence, which Rothstein is always sure to demand of those who don't agree with him. Besides, most of the serious research points in the opposite direction. For example, the University of Chicago's Susan Mayer undertook a far more comprehensive analysis of the relationship of income to school achievement in a 1997 book, *What Money Can't Buy*. She examined nearly 17,000 records in two massive data sets in her search for the true effects of income. The study is important for its methodological sophistication and its conclusions, which take us beyond the traditional left-right political axis regarding welfare programs and the causes of poverty.

While we await socioeconomic utopia, Rothstein would expect nothing better from the schools themselves.

Mayer showed that income per se is not a consequential factor in children's performance. Beyond providing the ability to satisfy basic needs like food and shelter, income is not a necessary, much less a sufficient, explanation of children's academic achievement. Mayer found that a supportive family structure (a stable, two-parent home), a culture of learning within the family and neighborhood, and natural abilities were much more important than income. Given these factors, income can certainly help people achieve their ends. In their absence, however, income is largely inconsequential.

These findings would seem to present the perfect opportunity for Rothstein to flex his critical muscles by rebutting Mayer's scholarship—or at least addressing it. Nevertheless, he seems to have ignored her study entirely—despite the fact that her findings run directly counter to his Big Idea.

What's strange is that Rothstein certainly isn't shy about attacking his ideological opponents. In his first column of 2001, he tackled a much-publicized Heritage Foundation study of 21 schools that exhibit both high poverty rates and high test scores. This was a direct challenge to Rothstein's Big Idea—that there is an immutably inverse relationship between income and student achievement. How would Rothstein assail the study? Let's try invoking ideology first. Heritage is "conservative," we read, so its report "as a whole is enveloped in such contempt for most public education that its valid messages are lost." Then let's supplement the Big Idea. "The report's biggest flaw is its assumption that poverty alone defines the problems of the low-income school.... Rather, schools with consistently low scores typically have children for whom poverty per se is only one problem. They also suffer from crime-ridden neighborhoods, broken families, parental stress, inadequate housing, and poor health."

Rothstein's maneuver here is a tactical retreat. Heretofore, income has served for him like an engine's governor, an upward limit on a school's (or individual's) capacity to perform. But Heritage located schools that perform well despite their low-income student body. Rothstein's way out is to declare the study myopic, saying that it focuses exclusively on income, when other criteria, not examined in the study, are clearly the relevant factors. However, this maneuvering is a bit circular, considering that inadequate housing, bad neighborhoods, and poor health are usually direct proxies for poverty. Still, in an effort to avoid a contradiction, Rothstein makes an important acknowledgment: successful children, he says, come from "stable homes and parents [who are] regularly employed." Hence, Rothstein suggests, if only the study had asked proper questions, instead of obsessing over income, it would have found those factors (like parents) that really contributed to the schools' academic success. Here it sounds as if he agrees with Mayer—when it suits his ideological purposes.

The Bell Tolls

So fixated is Rothstein on the determinative role of income that he actually proposes a sliding scale for learning standards. It is inherently unfair, he has argued on several occasions, to judge all schools and students by the same academic standards. Because of their disadvantages, low-income students should not be expected to compete against the affluent. He recommends adjusting the expectations for the schools located within a given neighborhood based on the area's average income.

Rothstein broached this argument in a November 1999 column entitled, "Does Social Class Matter in School?" We are in no suspense about Rothstein's answer, of course: the poor, because they are poor, should not be expected to learn as much. Poor children are stuck in shoddy day care, while "typical middle class parents raise their children differently," providing them their own "head start." As Rothstein puts it, we must resist the "dangerous myth" that "all children, regardless of academic background, can achieve to the same high standards if only schools demand it." Accordingly, "schools with privileged children should be termed failing if they test only at the 70th percentile.... A school with many poor children, scoring at the 35th percentile, could be highly successful, though it tests below average." Substituting race for class in such an argument would make it perilously near to the views that were attributed to Charles Murray and Richard Hernnstein in their controversial book *The Bell Curve*.

Rothstein actually invoked Murray's controversial theories in a December 1999 column, presumably because he found this framework useful: "We simply cannot set one standard applicable to all," he writes, since that would be "statistically foolish." The foolishness derives from the observation that human performance is often normally distributed, which means that some are necessarily one or more standard deviations above or below the average. It follows for Rothstein that differing income levels call for differing academic standards.

Most schools do something like this by "tracking" students, albeit by previous academic performance, not income. Rothstein, however, denounces this form of relativism, apparently because it is not income-based. For him, tracking is based on another "dangerous myth," since the practice of taking "poor children . . . assigned to tracks and taught less challenging lessons" assumes that "disadvantaged children could not learn." So you can offer all students, rich or poor, the same curriculum; you just can't test to see whether they've all mastered it. Evidently the only solution is a classless society.

The danger is that, while we await socioeconomic utopia, Rothstein would expect nothing better from the schools themselves. Indeed, his answers to the problem of disparities in student achievement often involve placing limits on high-achieving children rather than improving the education of low-achieving children. Consider homework. In a May 2001 column, Rothstein lamented the fact that teachers are assigning more homework, which is said to be "up 50 percent in the last two decades." This is a problem, Rothstein believes, because it "may increase the gap between students from middle-class and low-income homes. With growing inequality now a greater danger than middle-class pupils' inadequate achievement, policies that widen learning differences should be avoided."

Rothstein cites an academic authority to reinforce the claim, quoting University of North Carolina professor Eugene Brooks, who says, "Because of homework, schools either consciously or unconsciously reproduce social inequality. It can be avoided only if teachers take over homework supervision from parents." That's a somewhat breathtaking mission for the school—reducing the impact of social class on learning by expunging parents from the equation, since they are unequal in their degree of helpfulness. It is apparently better for all youngsters to languish in dreary study halls—presumably reducing the amount of time left for instruction—than to take the risk that one mother might help her child learn faster than another. Rothstein concludes, "It is unconscionable for educators to exacerbate inequality by assigning homework without first ensuring [that social programs] are in place." Perhaps Rothstein ought to teach 5th grade. No homework, no testing, and being graded on their "zest for living" rather than on their achievement in math and science—his students would love him!

Hungry for Facts

Though Rothstein's column is nominally about education, regular readers find themselves spending as much time rooting about in sundry social programs, organized around the theme that those who would fix schools by attending directly to schooling are misguided. "By focusing only on schools, government may waste money trying to fix academic problems that it could have prevented in the first place at less expense." For example, if schoolchildren falter on tests, the reason must be malnutrition. In a January 2000 column, he writes, "Forced to spend more, poor families often raid food budgets to pay rent. Children then suffer nutritionally, compounding cognitive problems."

New York Times

Rothstein ought to teach 5th grade. No homework, no testing, and being graded on their "zest for living"—his students would love him!

He amplified this theme in August 2001, asking, "What is the most efficient way to raise low-income pupils' achievement? . . . Improving nutrition might bring a bigger test-score gain. . . . Undernutrition found in the United States affects performance. . . . Reducing hunger that causes low test scores may be accomplished more easily than the various unproved educational reforms commonly advanced. A higher minimum wage that helps low-wage parents feed their children could be one important step in ensuring that no child is left behind. And those wanting to narrow academic gaps can . . . contribute to the nation's overstressed food banks. This too would amount to real education reform."

A claim that malnutrition might affect cognition and learning is not unreasonable. But there is little credible evidence that a large hunger problem exists in America. Activist groups that routinely assert that American children suffer malnutrition are actually measuring something they term "food insecurity," a condition that usually results more from poor household management than from scarce resources.

Rothstein has shown himself to be as gullible to the claims of activists as the popular media. When a 1990s survey was said to show that there were 12 million hungry children, NBC's Tom Brokaw quickly transformed that claim into "12 million American children are malnourished." On CBS, Dan Rather went further, announcing, "A startling number of American children [are] in danger of starvation." These preposterous facts were generated by the Food Research Action Center (FRAC) and rely on some statistical sleight of hand. The surveys don't measure nutrition, but instead ask about feelings of "insecurity" concerning food, measuring the percentage of households that reported "difficulty getting enough nutritious, safe food at all times, in a socially acceptable way." This approach has drawn scorn from such commentators as the *New Republic's* Mickey Kaus and Johns Hopkins professor George Graham, who said of these numbers, "A lot of what the activists are calling hunger is just absolute rubbish . . . irresponsible people making irresponsible claims." Kaus was more succinct: "The whole project oozes phoniness."

In fact, the larger food-related problem besetting American children today is not hunger but obesity. Indeed, one reads in one of Rothstein's November 2000 columns, here citing the U.S. Surgeon General, that "with the percentage of overweight children doubling in the last 30 years, school kids are less healthy."

In the obesity column, Rothstein is defending mandatory phys ed, which he worries could be sacrificed to increased academic demands. He wonders, "Is all this (obesity) an unintended consequence of raising academic standards? Are schools squeezing in more math by eliminating gym?" This was not the first time he made this claim. In a November 1999 column, Rothstein, fretting that schools are being carried away by standards and testing, wrote, "Test scores are rising. But more teenagers have fatty diets. . . . adolescent obesity has soared. Neither children nor our economy will prosper if graduates have higher scores, but are overweight." Surely Rothstein can make better arguments against testing than the shaky hypothesis that it leads to obesity. Are testing programs really taking the place of physical education? Do existing phys ed programs make any difference in children's weight? Here Rothstein is engaging in pure conjecture.

New York Tim

Rothstein's assertions are rarely backed by solid research or even by decent facts.

Continuing his seeming obsession with the relationship between food and cognition, Rothstein writes, in a March 2001 column, "Surgeon General David Satcher reported . . . that more than a third of poor children have untreated dental cavities. Pupils taking tests with toothaches are unlikely to score as well as those undistracted by pain." In other words, we're unfairly expecting teachers and schools to administer tests before society has ensured adequate dental care for everyone. But are most cavities even associated with toothaches? Does Rothstein have any evidence that toothaches impair performance? And with all this social engineering, when will we ever teach reading?

Double Standards

Rothstein's own assertions are rarely backed by solid research or even by decent facts. However, when he is confronted with research that counters his worldview, the methodological gloves come off. Consider his approach to voucher research. If poor families were given vouchers redeemable at the schools of their choice, and the achievement of some students rose, it would call into question Rothstein's notion that income is the master variable. We therefore ought to expect him to resist such research. And that is exactly what he has done.

Confronted with studies that show black students making gains as a result of being given vouchers (full disclosure: these studies were led mainly by Paul Peterson, this journal's editor-in-chief), Rothstein turns to Stanford education professor Martin Carnoy to build a strict critique. Proper experiments should be double-blind, Carnoy says, as in drug trials. This cannot be done with kids who know what school they're in. Moreover, some students left the voucher program after one year, and some declined to participate at all. Rothstein writes: "Researchers cannot know if leavers differ significantly from stayers." The gains reported were "small" and "inconsistent."

There are many problems with this critique. A study doesn't need to be double-blind in order to be valid; many studies in the medical field are not double-blind. Moreover, the researchers used well-regarded statistical techniques to control for the fact that some students left the program after one year or didn't participate at all. The broader problem is that Rothstein fails to use this standard, or anything remotely like it, when tackling research that favors his Big Idea. The information that he routinely asserts with confidence, such as studies supporting the primacy of economic factors, represents inferences based on, if anything, methodological procedures that were far less rigorous than those used in empirical work on vouchers. The impression offered is of one straining at gnats while swallowing camels.

There is more, much more. The pattern persists from week to week. The Big Idea is defended at all costs. Critics are dismissed. Economics and social class explain just about everything—except where they don't. America's self-styled newspaper of record has but one regular commentary on education, and its author is gripped by a worldview that often blinds him to evidence and confounds the rules of logic.

–David W. Murray is the former director of the Washington-based Statistical Assessment Service, a nonpartisan organization that examines the use of quantitative research by the media.

essay review

by WILLIAM G. HOWELL

Data Vacuum

Only larger voucher experiments will yield answers

School Vouchers: Examining the Evidence
By Martin Carnoy
Economic Policy Institute, 2002.

Rhetoric Versus Reality: What We Know and What We Need to Know About Vouchers and Charter Schools
By Brian Gill, P. Michael Timpane, Karen Ross, and Dominic Brewer
RAND Corporation, 2001.

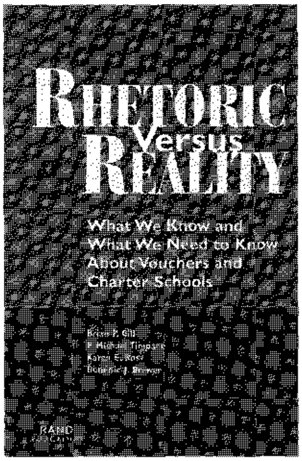
School Vouchers: Publicly Funded Programs in Cleveland and Milwaukee
General Accounting Office, 2001.

"The Hidden Research Consensus for School Choice"
By Jay P. Greene, in *Charters, Vouchers, and Public Education*, edited by Paul Peterson and David Campbell
Brookings Institution Press, 2001.

"Market-Based Reforms in Urban Education"
By Helen F. Ladd
Paper presented at the Seminar on Creating Change in Urban Public Education, December 7–8, 2001, Cambridge, Massachusetts.

"Vouchers for Elementary and Secondary Education"
By Isabel Sawhill and Shannon Smith, in *Vouchers and the Provision of Public Services*, edited by Eugene Steuerle, Van Doorn Ooms, George Peterson, and Robert Reischauer
Brookings Institution Press, 2000.

"What Research Can Tell Policymakers about School Choice"
By Paul Teske and Mark Schneider
Journal of Policy Analysis & Management, 2001.


Rhetoric Versus Reality
What We Know and What We Need to Know About Vouchers and Charter Schools
RAND

In their sheer volume, reviews of the small body of empirical work on school vouchers are beginning to eclipse the research literature itself. It is not often that so much is written about so little, but scholars are jostling to have the final word on what we know about vouchers as the Supreme Court prepares to rule on their constitutionality.

All of the reviewers decry the contentious political and ideological haranguing that surrounds the public debate over vouchers, and many blame the media for conspiring in the frenzy. As a corrective, each promises a clear-headed, impartial account of the empirical findings on school choice. Given the expressed motives of the reviews and the modest size of their subject matter, one would hope that a consensus would emerge.

Would that it were so. The authors find little about which to agree. Only in the face of overwhelming and unambiguous evidence does any consensus emerge. When scholars survey the nascent empirical literature on school choice, they see very different things and discern very different lessons. And they will continue to do so until more, and better, data are collected from larger, better-financed voucher programs.

The Reviews

Stanford University education professor Martin Carnoy declares his intent to interject a "balanced perspective" in an empirical literature written "mainly by researchers who openly and actively support vouchers" and media that "report results from these analyses without necessary caveats and alternative views." Carnoy is deeply skeptical about the positive test-score impacts observed for African-Americans in the randomized field trials conducted in New York City; Dayton, Ohio; and Washington, D.C. He concludes: "The question to ask is not whether these latest Peterson-group reports overestimate private school effects, but by how much." (Full disclosure: Paul Peterson, editor-in-chief of *Education Next*, and I were among the scholars conducting these evaluations.) Carnoy is no less critical of Jay Greene's recent analysis of the Florida voucher program—or what Carnoy calls "the latest round of voucher advocacy research." (See Letters, p. 7, for the debate between Greene and Carnoy. See "The Looming Shadow," *Research*, Winter 2001, for Greene's study of the Florida A+ program.)

Like Carnoy, Helen Ladd argues that the evidence on achievement is at best "preliminary," and if it supports the claims of voucher advocates, it does so only under very restrictive conditions and for a very small subset of the urban poor. However, the Duke University economist is less distracted than Carnoy by the supposed misrepresentations of voucher advocates posing as researchers and more impressed by an empirical literature that consistently turns up negative findings. Ladd argues, for instance: "The evidence simply does not support the claims of those who argue in favor of more parental choice and competition on the instrumental grounds that it will make an education system significantly more productive than it would otherwise be." Though the only data on large-scale voucher programs she finds convincing come from her own study of the New Zealand public-choice program and Martin Carnoy's work in Chile, Ladd confidently asserts that vouchers will not help those public schools that are having the hardest time of it. "The bottom line is clear: Large-scale expansion of parental choice and competition will not, by itself, solve the problems of the most distressed urban schools." If anything, the effects of vouchers on the schools that are "left behind" are surely "adverse."

Much like Carnoy's and Ladd's reviews, RAND's book-length summary of the literature emphasizes the limitations of the existing empirical work on school choice. Its general assessment of the findings, however, is considerably more positive. According to Brian Gill, P. Michael Timpane, Karen Ross, and Dominic Brewer, the book's authors, all of the empirical evidence supports the contention that vouchers improve parental satisfaction. The authors say that the existing research on student achievement, while still preliminary, bodes well for vouchers. "Small-scale, experimental privately funded voucher programs targeted to low-income students suggest a possible (but as yet uncertain) modest achievement benefit for African-American students after one or two years in voucher schools." Contrary to the claims that voucher programs "cream" the best students from public schools, RAND's scholars find that voucher programs have successfully placed "low-income, low-achieving, and minority students in voucher schools." They raise concerns, however, about the underrepresentation of students with disabilities and poorly educated parents in private schools. They also find that while targeted voucher programs may "modestly" alleviate racial segregation patterns in urban cities, the evidence from large-scale, unregulated choice programs suggests that vouchers may increase levels of social stratification.

In August 2001, the General Accounting Office issued a tepid report on the publicly funded school voucher programs in Cleveland and Milwaukee. (For reasons that are never fully explained, the report ignores all evidence from privately financed programs in the United States as well as all choice initiatives conducted abroad.) Both the Cleveland and Milwaukee programs successfully targeted poor, lower performing, and predominantly minority students in public schools. On average, voucher students were placed in smaller classes in smaller schools with less experienced teachers. The GAO report describes the research on student achievement as, at best, mixed and preliminary. While some "contract researchers found little or no significant improvement in voucher students' achievement," other investigators "found some positive effects." The report's conclusion laments the generally poor quality of the existing studies and advocates state funding for the collection of more achievement data and the conduct of studies that yield higher response rates.

When scholars survey the nascent empirical literature on school choice, they see very different things and discern very different lessons.

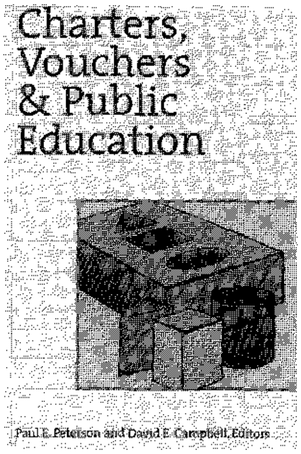
Brookings Institution scholars Isabel Sawhill and Shannon Smith give the existing literature on school vouchers a cautiously positive review. Virtually all studies on school vouchers, Sawhill and Smith write, show that private school parents are markedly more satisfied with their children's school than public school parents. But after reviewing the reports issued by a trio of research teams studying the Milwaukee program—one headed by the University of Wisconsin's John Witte, another by Harvard University's Paul Peterson, and a third by Princeton University's Cecilia Rouse—Sawhill and Smith claim that "it is simply not possible at the current time to render a clear verdict on the outcome of the experiment." Finally, Sawhill and Smith anticipate that vouchers may lead to "a greater segmenting of recipients by race, ethnicity, income, or ability," and hence recommend that further research be conducted on the subject.

The reading of the literature by Paul Teske and Mark Schneider, political scientists at the State University of New York at Stony Brook, is more optimistic still. Teske and Schneider note that the existing empirical work on school vouchers is quite positive on a variety of issues: academic considerations appear paramount when parents choose schools; voucher recipients are more satisfied with their schools than their peers within public schools; and vouchers lead to "clear performance gains for some groups of students using the vouchers, particularly blacks, compared with the control group." Teske and Schneider caution policymakers, however, that a rising tide of evidence suggests that school vouchers may lead to greater social stratification and racial segregation.

Finally, Jay Greene sees "uniformly positive" findings in the school voucher studies. While "one would never know it from the media coverage," there exists a "hidden consensus" among scholars that vouchers work. Students who use vouchers score significantly higher on test scores than their public school peers—just as they are more tolerant and their parents are more satisfied. Though enterprising parents among the disadvantaged are more able to avail themselves of choice than others, Greene observes that this is no less true of "food stamps, Temporary Assistance for Needy Families, and virtually all other antipoverty programs." He notes that, although few studies have examined the impact of choice on public school students, most every finding to date suggests that vouchers, rather than adversely affecting students who are "left behind" in public schools, actually lead to gains for public and private school students. Compared with other education interventions, Greene concludes, "school choice has been thoroughly and carefully studied" and has generated "consistently positive results."

This sweep, from Carnoy to Greene, is hardly exhaustive. Indeed, the reviews keep coming. Paul Hill and Tom Loveless, both senior scholars at the Brookings Institution, recently acquired a million-dollar grant from the Bill and Melinda Gates Foundation to assemble a panel of experts on school choice, with the goal of writing the definitive statement on what we know and what requires additional study. The National Working Commission on Choice in K–12 Education, as it is called, intends to trump all of the reviewers listed above. Stay tuned: two years from now, the commission will release its conclusions.

For now, though, let's look at what these reviews add up to. On one issue there is general agreement: parents who use vouchers are more satisfied with their private schools than are parents who apply to voucher programs but remain in public schools. Everywhere else, however, the reviews turn up inconsistencies and disagreements. While RAND, Teske, Schneider, and Greene suggest that the best evidence on student achievement comes from the new randomized field trials, what some call the "gold standard" of social-scientific research, Carnoy and Ladd remain fixated on two sources of potential contamination—attrition rates and response bias—and are inclined to dismiss (or at least heavily discount) the findings. While Teske and Schneider argue that parents are principally motivated by academic concerns when choosing schools, Ladd sees considerable evidence that parents select schools on the basis of their ethnic composition. Though RAND concludes that "nothing is yet known empirically about the civic socialization effects of voucher and charter schools," Greene lists a batch of studies (some of which were published after the RAND report went to press) that consistently show that private schools are doing a better job of teaching tolerance. While many of the reviews anticipate the possibility of social stratification in large-scale voucher programs, Ladd sees it as a foregone conclusion.


Charters, Vouchers & Public Education
Paul E. Peterson and David E. Campbell, Editors

The reviewers disagree about the content of the empirical findings, their implications for different kinds of choice initiatives, and the quality of alternative research designs. They disagree, if only implicitly, about the relevance of choice experiments conducted abroad. They disagree about which issues require further study and which have been settled conclusively. In searching for consensus, these reviews only underscore how much disagreement lingers.

The Chimera of "Objectivity"

Lawrence Hedges of the University of Chicago, in a recent report on education research, identified the basic problem facing policymakers who are considering choice initiatives. "To put it bluntly," Hedges wrote, "in the face of diverse and apparently conflicting information about the potential effects of a policy, it is difficult to know which studies to believe." Each of the reviews summarized above tried to rectify this situation. Unfortunately, disagreements among the reviewers only replicate controversies among primary researchers, and the clouds have yet to part.

All of the reviewers recognize a need to inject a measure of reason into the politically charged debate over school vouchers. RAND puts it this way: if policymakers are to make an informed judgment on vouchers, they need "a thorough and objective empirical assessment" on the literature. RAND promises "accurate data," "careful objective analysis," and "a clear picture of the choices [we] face in educating America's citizens." RAND is quite sure that it has it right. The trouble, though, is that all of the other reviewers think they have it right as well. Given the manifest contradictions across the reviews, the defense of impartiality begins to break down.

Rhetorically, it is perfectly obvious why all of these reviews stake a claim to objectivity. All hope to distance themselves from the ideological warring between die-hard advocates and committed critics of school choice. All want to gain traction in a debate that has become increasingly politicized, and the obvious way to do so is to appeal to reason; to plead with the reader that if he would just set aside his ideological precommitments, he would see exactly what the reviewer sees. Then, with the reviewer's credentials in order, and the reader's trust secured, genuine progress can be made.

Such claims to objectivity, however, are misplaced. In their attempts to provide neutral accounts of the empirical literature on school choice, all of these reviewers necessarily fail. They fail not because of their individual flaws—though some reviews certainly are more defensible than others. They fail because the act of assigning meaning to facts is unavoidably subjective, especially when the facts come from such a small empirical literature. While the laws of science guide the practice of research, they cannot say what constitutes "persuasive" findings. They cannot specify the threshold of evidence required to "prove" a hypothesis. They cannot determine which facts have important policy implications and which are less consequential. As the education philosopher Francis Schrag notes, when evaluating (as opposed to doing) empirical research, "normative considerations appear to be decisive in determining the admissibility of facts and in providing a lens through which these facts are filtered."

Each reviewer is correct to note that the others offer selective reviews of the evidence, adjust the standards of evidence according to the substance of the findings, and find guidance in their ideologies. Each, however, is wrong to exempt him or herself from such charges. All have strong normative commitments—about the purposes of public education, the promise and failings of markets, the appropriate level of government regulations—that inform their assessments of the modest empirical literature on school choice. To make sense of the findings within this literature, the reviewers invariably draw on these normative commitments. As sociologist Joel Best writes in *Damned Lies and Statistics*, "We may think of statistics as facts, but people make facts meaningful, and analysts' ideologies shape the meanings they assign to social statistics."

Scholars' reassurances that they have their intellectual houses in order and can clearly see the empirical evidence on school choice as it is, and for what it is, amount to very little. Indeed, there is a certain irony that in attempting to distinguish truth from fiction and to distill the values and speculations and ideologies that apparently muddy the public debate on school vouchers, these reviews, when read collectively, only make matters worse.

Their subjectivity, of course, does not disqualify any of these scholars from continuing to do primary research on education. Should scientific progress depend on scientists' hearts being pure, progress most assuredly would be illusory. As long as scholars' findings are subject to the scrutiny of others—what the philosopher Karl Popper identifies as the requirements of "intersubjective testability"—scientific inquiry can proceed amid deeply held ideological commitments.

Advancing Knowledge

There is nothing wrong with hearing scholars' views on the state of the empirical literature on school choice. I myself found them helpful in trying to organize my thinking about the existing body of research. Knowing their views, however, should not be confused with knowing more about the likely consequences of various kinds of school voucher programs. If we want to advance knowledge, researchers need to confront two challenges.

First, we simply need more data. Until the early 1990s, the only available evidence on school choice came from programs that limited alternatives to schools within the public sector. In the mid-1990s, a highly contentious debate erupted over whether a couple of hundred voucher students in several private schools in Milwaukee performed higher on standardized tests than their peers in local public schools. Without any solid data to anchor it, the public debate on vouchers floated adrift.

To be sure, the state of research has improved. New evidence has been collected from publicly funded voucher programs in Florida and Cleveland; new findings from randomized field trials conducted in New York City, Dayton, Charlotte, and Washington, D.C., are now available; and a growing body of evidence collected abroad provides a comparative perspective on the probable effects of large-scale choice initiatives.

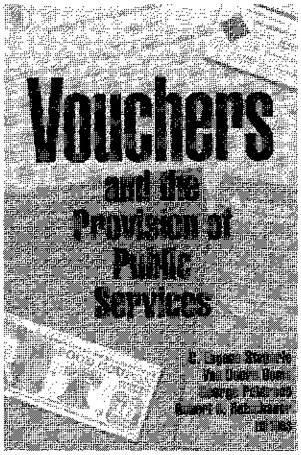
Still, we lack the data required to address a host of critical questions about school choice. RAND correctly notes that "the list of unknowns remains substantially longer than the list of knowns." We know very little about the long-term consequences of vouchers on student achievement. Little or no data have been collected on the effect of school choice on graduation rates, incarceration rates, the probability that students will end up on welfare, the chances that they will be employed full time—all outcomes that deserve careful scrutiny. For the simple reason that all of the existing domestic choice initiatives are quite small, we still know very little about the systemic effects of vouchers on existing public and private schools.

The only way to answer these questions is to begin collecting data from larger programs and investing more money in evaluations so that students can be tracked for longer periods of time. The need for richer data constitutes a primary reason why fellow researchers and I support the creation of a citywide program that offers vouchers set at the amount of per-pupil funding in area public schools. Until quality data are collected from larger voucher programs, ideology will continue to serve as a cheap substitute for hard evidence on the likely effects of school vouchers.

Some will argue that supporting additional research in itself represents a form of advocacy and cannot be justified on purely objective grounds. At one level, this undoubtedly is true. Calling for more research ultimately rests on an intuition that vouchers may achieve important public objectives and that studying them will reap valuable information about the comparative strengths of public and private education systems. Given that no study has demonstrated that targeted urban voucher programs hurt students, and several studies have shown that they are especially beneficial to low-income African Americans, I am perfectly comfortable making the call. It is worth noting, too, that insisting on an end to research also is a form of advocacy. There is nothing neutral about arguments that we should suspend efforts to study the effects of school choice initiatives.

The second challenge researchers face concerns the statistical techniques and measurement tools currently available to evaluate educational policy. Consider, for example, the problem of self-selection bias that plagued the literature of school sector effects. While impressive data sets were assembled on the achievement levels of thousands of students in public and private schools, statisticians could not be sure whether observed differences reflected the quality of the schools or the students who self-selected into them. It was not until the 1990s that education researchers began to employ a research design that effectively mitigated the problem of self-selection. By using a lottery to assign vouchers, researchers assembled treatment and control groups that, at baseline, were indistinguishable from one another. Subsequent differences observed between the two groups could then be attributed to the intervention, without the need for estimating complex statistical models. The introduction of randomized field trials to education research is as much a boon to knowledge as the results from any single study on school choice.

Unfortunately, plenty of statistical problems besides selection biases continue to wreak havoc on education research. A series of excellent papers by economists Thomas Kane, Douglas Staiger, and Dale Ballou (see "Randomly Accountable," *Education Next*, Spring 2002, and "Sizing Up Value-Added Assessment," this issue) scrutinize the error built into value-added test-score measures, many of which are used in state accountability systems. The imprecision of statistical models that estimate year-to-year changes in student test scores to evaluate the quality of individual schools and teachers is sufficiently large that accountability systems frequently sanction success and reward failure. When virtually all education interventions yield rather modest test-score changes from year to year, it becomes extremely difficult to detect effects given the amount of statistical noise in our instruments. The problem of measurement error is quite real. Until we find ways of mitigating its influence, our ability to judge the effectiveness of education policies reliably is limited.



If scholars are genuinely interested in making advances, they ought to worry less about assembling panels of experts with an appropriate balance of critics and advocates and focus instead on research design issues and, whenever possible, support the collection of more and better data.

Conclusion

The evidence collected from a long-term, well-financed voucher initiative may not change many minds, least of all those already committed to one side of the debate. The findings, however, will help set boundaries on what constitutes a reasonable argument either for or against choice. The forecasts of widespread innovation brought on by competition and of cultural balkanization by privatization will have to accommodate an empirical reality that is considerably more nuanced than vouchers' most ardent advocates and critics are currently willing to admit.

Our reviewers generally agree about one issue (the greater levels of satisfaction among voucher recipients) not because they agree ideologically or because their assessments of the research on this topic are principled and objective. Instead, they agree because they have no other choice. The existing evidence on the topic overwhelms any possible normative objection. Scholars continue to disagree about other topics (such as the likely impact of a large-scale voucher program on the education of students who remain in public schools) because the evidence assembled to date is more provisional.

The proper antidote to ideology is not an insistence of objectivity, but more and better social-science research. Pious sentiments do not make for objectivity; careful research does. In medical research, many scientists believe in the product they are studying—why else would they work on it? Yet the findings from randomized field trials discipline their thinking and restrict the conclusions they can credibly make. The once dominant views that nicotine is not addictive, or that sodium intake is universally detrimental to one's health, or that cholesterol buildup is the primary cause of heart disease have all been rebuffed, not for lack of powerful interests working on their behalf, but because a mountain of empirical evidence discredits them.

None of the current participants in the school voucher debate will cast the decisive vote on the meaning of the existing research. The next generation of researchers will, and currently they are enrolled in the private and public elementary schools being studied. The best way to help them is to develop new methods for evaluating education outcomes and to assemble high-quality evidence on the long-term programmatic effects of different programs, serving different populations, in different geographic settings.

–William G. Howell is an assistant professor of political science at the University of Wisconsin–Madison.

Surface Wounds

Thus far, public schools have hardly noticed competition

Revolution at the Margins: The Impact of Competition on Urban School Systems

By Frederick M. Hess

Brookings Institution, 2002, $45.95; 268 pages.

As reviewed by Edward B. Fiske

REVOLUTION
at the Margins
THE IMPACT OF COMPETITION ON URBAN SCHOOL SYSTEMS
Frederick M. Hess

For the most part, the language of economics has informed the public debate over school choice. Free-market economist Milton Friedman was the first to develop the concept of school vouchers, and some of the most enthusiastic supporters of charter schools and vouchers are businessmen who view the delivery of public education through the lens of their own experiences in creating and marketing goods and services. It is thus important to know that Frederick Hess is a political scientist, not an economist. What interests him is the way in which the political and organizational realities of urban schools influence their responses to competition and thus help determine how competition will affect the schools that more than 90 percent of students still attend.

The underlying agendas of choice advocates vary widely. Some view school choice as a social good in and of itself, while others may have indirect objectives, such as funneling public funds to religious schools or privatizing public education. Whatever their agendas, however, most supporters of school choice build their political case on the virtues of competition for public education as a whole. "Perhaps the most commonly advanced argument for school choice," writes Hess, "is the notion that markets will force the nation's public schools to improve, particularly in those urban areas where improvement has proved so elusive." It is this argument—that a rising tide of choice and competition will raise all boats—that Hess explores in this thoughtful volume.

Hess offers a litany of ways in which school systems differ from private enterprises. Whereas a business can focus mainly on serving its customers, public schools serve a multiplicity of stakeholders, of which parents and students are only two. Also, Hess observes that educators tend to be driven by intrinsic motives, such as a "sense of calling," that lessen the ability of supervisors to force them into a competitive mode. Losing students is not always a threat to public school teachers and managers. In fact, it can be downright attractive when enrollments are rising or when choice relieves them of disgruntled parents or low-performing students. When competition becomes bothersome, it can be handled with political responses, such as running advertising campaigns or creating new high-profile programs, that scarcely relate to educational performance. Moreover, school systems tend to reward the following of rules and procedures, which makes life difficult for those with an entrepreneurial spirit. Finally, education is an "unwieldy market good," where it is difficult to define quality—or even to know it when you see it.

Losing students is not always a threat to public school teachers and managers.

This discussion is informed by Hess's case studies of three cities where the conflict of economic theory and political-organizational reality could be observed firsthand. He concedes that these case studies are not intended to be definitive, suggesting that they may be thought of as "theoretically directed journalism" rather than "conventional social-science scholarship."

Hess devotes two quite thorough chapters to describing the decade-long voucher experiment in Milwaukee. There, choice did not compel the school system to change its routines, but it did lead to a loosening of "bureaucratic procedures and organizational routines" that allowed some educational entrepreneurs to do their thing. Hess calls this the "pickax" response to choice—poking some holes in the system but not making any major changes.

Hess's second case study focuses on the five-year-old voucher plan in Cleveland, where he finds that the potential benefits of choice and competition were neutralized by multiple factors, including frequent changes in leadership, the state's move to take over the city's schools, the modest size of the vouchers (only $2,250), and the existence of strong unions. As a result, he concludes that school officials saw vouchers as "a largely symbolic threat" and felt no need to respond.

Finally, Hess examines the three-year-old Horizon scholarship program in the Edgewood Independent School District of San Antonio. The program offers privately financed scholarships to low-income Hispanic students to attend private schools. Edgewood shows, Hess writes, "how the districts may respond in ways that have little, if anything, to do with educational quality," including a public-relations campaign, a management study, and the opening of an already planned Fine Arts Academy.

Hess uses these case studies to speculate on how choice might be introduced in ways that both respect the built-in political and organizational constraints of urban school districts and lead to school improvement. "The lesson is not that markets cannot drive more profound change in education," he writes, "but that such effects will require changing the institutional and organizational context of urban schooling.... In short, making competition work as intended will require much more than the simple introduction of market mechanisms."

Hess questions the pickax approach as a half-measure that cannot compel change or "alter the incentives that drive educational performance." He likewise questions the workability and desirability of the alternative approach of "unchaining the bulldozer," or moving to a ruthless competitive system oriented toward coercive accountability and other extrinsic incentives. (I find the metaphor a bit curious, since I have trouble picturing a bulldozer—not a bulldog—on a chain.)

On the matter of starting new schools, Hess suggests that, given the financial and other obstacles involved, the most likely source of choice entrepreneurs is people "operating out of religious obligation or a desire to reap a profit." This observation fits the current scene. Channeling public funds to parochial schools was reportedly part of the political deal that established the Cleveland voucher plan (an observation that Hess curiously underplays in his narrative), and two-thirds of charter schools in Michigan are now run by education management organizations such as the National Heritage Academies.

Hess states at the outset that he is "not seeking to provide a definitive account of educational markets, but to launch a more useful conversation on the topic," and he has achieved this goal. Economists have always understood that a variety of factors can interfere with the smooth functioning of markets, and Hess reminds us that the same principle applies to education markets as well.

Districts may respond in ways that have little, if anything, to do with educational quality.

Hess succeeds in posing a challenge to those who see choice and competition—the manipulation of incentives, if you will—as a way of improving schools without getting bogged down in the nitty-gritty issues of providing a quality education. "In fact," Hess concludes, "educational competition cannot be divorced from discussions about testing, teacher certification, school district governance, educational administration, or other frustrating conversations that many school choice proponents have long wished to avoid. In the end, the fate of educational markets, for good or ill, is intertwined with broader issues of educational politics and policy."

–Edward B. Fiske, a former education editor at the New York Times, *is co-author, with Helen F. Ladd, of* When Schools Compete: A Cautionary Tale *(Brookings, 2000).*

As reviewed by John Gardner

As a member of the Milwaukee school board, I am one of the officials responsible for the school district's response to the city's voucher program. I have helped to manage the changes necessary to make the district's schools educationally competitive, and I have witnessed the ebb and flow of competitive responses and reactive managerial and political opposition. The Milwaukee school choice program and the response of Milwaukee Public Schools are especially significant in light of Frederick M. Hess's study of the effects of competition on large urban school districts. One of Hess's conclusions is that among the three American cities where vouchers have been tried, only Milwaukee has a program with the size, duration, and per-pupil payments to test the hypothesis that market competition can improve schools.

Hess's other case studies include the state-funded voucher program in Cleveland and a privately funded scholarship program in the Edgewood district of San Antonio, Texas. He argues that the choice programs in Edgewood and Cleveland never grew large enough for the local school districts to really notice any loss of students or funding. The publicly funded Milwaukee Parental Choice Program, by contrast, had grown to almost 7,000 students by the 1998–99 school year, the final year of Hess's study. This was at least a sign that a substantial number of students and parents were unhappy with the quality of the district's schools. Yet Hess finds that even in Milwaukee the response of the school district was focused on buffing its image much more than on actually improving the system.

In my opinion, Hess's only limitation is his stopping point, the end of the 1999 school year. Up until then he has the story right. He argues that the leadership of the Milwaukee schools, the school board and the central office, attempted to respond with systemic and school-based reforms, but were perpetually thwarted by managerial and political constraints that he describes with chilling insight and accuracy. They—we—managed to liberate a handful of schools from managerial, contractual, and bureaucratic constraints that had stifled improvement and reform. Expanding these reforms to all schools was constrained by political and managerial constraints at all levels. The board was unwilling to take on the powerful teacher union in many instances when it most mattered. There were not enough effective, entrepreneurial principals to turn around more schools. And the central office remained unresponsive and territorial, more threatened by innovative schools than supportive of them.

What the study's stopping point leaves out are the revolutionary changes that occurred from 1999 to 2001. Despite a partial counterrevolution, these changes have been sustained. Schools now control 95 percent of the district's funding, select their own teachers, and develop curricula to meet district and state standards. School budgets are based almost entirely on student enrollment, providing vigorous rewards and penalties for success or failure to attract students. The district's central office is reorganizing to serve, rather than control, Milwaukee's schools, in large part because the central office's budget is increasingly based on funds that schools choose to allocate to it.

Standards now include annual testing in basic subjects in grades 3 through 8. Students must now pass proficiency exams in order to enter and graduate from high school, replacing the system of social promotion. Incompetent teachers and administrators are now terminated through procedures instituted in, of all places, collective-bargaining agreements.

Before 1999, Milwaukee authorized one charter school, with only 70 students. Since then, the district has authorized 14, with more than 9,000 students attending. The number of previously oversubscribed specialty magnet schools has more than doubled. The number of popular K–8 schools has more than tripled. Before- and after-school childcare and youth recreation programs, once oddities within the system, are now in place in every elementary- and middle-school attendance area.

As a result of these and related changes, Milwaukee public school enrollment, parental satisfaction, funding, and academic achievement have all improved. Some of these changes may well have happened without the competitive challenge of school choice, but even choice's most vigorous opponents concede that at least some of the district's responses were produced by the threat of external competition.

Despite missing the revolution, Hess's analysis remains sound and moves the voucher debate helpfully away from the rigidities of the state-vs.-market debate. Education markets in large urban settings, he argues, differ so much from market analysis that one must question whether they are, in fact, even markets.

What Hess's stopping point leaves out are the revolutionary changes that occurred from 1999 to 2001.

Hess's dispassionate, balanced analysis will disappoint both opponents and advocates of market-based education reforms. Hess is hardly a defender of urban public education's failures. His previous book, *Spinning Wheels*, documented with depressing consistency the cycle by which urban districts adopt reforms, begin implementing them, encounter roadblocks, abandon reforms before their failure becomes apparent, and develop new reform agendas before the deadline to evaluate previous failures. In some districts the cycle has become a ritual procession of superintendent firings and curriculum changes, endlessly deflecting attention from managerial competence, school performance, and public accountability.

At the same time, *Revolution at the Margins* will offer little comfort to promoters of market-based education reforms who believe that markets, by themselves, are the answer. Milwaukee demonstrates that markets alone did little to change the public schools. It was the combination of a new market environment and effective responses from the public schools that simultaneously expanded choices for poor families and improved both choices and performance within the Milwaukee public schools.

What remains constant, whether urban districts respond effectively or respond at all, are the constraints and skewed semi-market within which they operate. School districts are, after all, governments, not profit-making enterprises. Like all governments, they respond to political pressures and public opinion more than to market signals and competitive incentives. School board members, central administrators, school principals, and teachers rarely have any significant stake in whether their schools lose some students. Even teachers, in some ways the most vulnerable to perceived or actual reductions in funding, are buffered from market pressures because of the national shortage of teachers, especially in major cities.

School officials and employees are, however, intensely interested in the funds that state legislators and municipal governments authorize. Rather than take on the difficult and challenging work of winning loyalty and enrollment from low-income parents, school districts face much more powerful incentives to fight legislative, legal, and public-relations wars. These offer the simultaneous advantages of familiarity, expectation, and shifting the blame for failure elsewhere.

Also diluting the competitive responses of school districts are the limitations placed on the choice programs themselves. Capping the number of students who can participate in choice programs has kept school districts from suffering any severe drops in enrollment. Similarly, in all three cities it was never clear that enough money left the school districts as a result of student defections to generate any significant economic distress among central administrators, principals, or teachers.

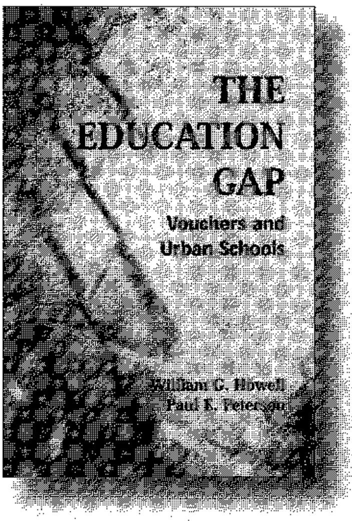
Finally, there is the thorny issue of

what the goal of competition is. Do schools exist to teach specific skills or to serve a broad range of social functions, including socialization, assimilation, acculturation, and the provision of a panoply of community services? Has a school successfully responded to competition if it drops its arts program in order to raise test scores in math? In this context, it's hard to tell who won the competition, or if the competition resulted in better outcomes for everyone.

Hess's most important contribution is clarifying and redefining the debate. School choice will ultimately prevail or disappear based on how it affects entire urban populations, not just the small group of students who benefit directly from being able to attend private schools tuition-free. Like all markets, public education operates within a framework of governmental mandates, regulations, and finance, both direct and indirect. Like all governments, public education ultimately depends on private market choices and very public political decisions. Hess persuasively shows that if markets are to have significant and positive effects on public education in major American cities, two difficult political revolutions must occur. First, markets must be created intentionally, not simply by introducing external competition, but by changing the internal rules and operations of the public district. Second, districts must develop the political will and managerial competence to respond—transformations that will, in most American districts, require their own revolutions.

School districts face powerful incentives to fight legislative and public-relations wars.

–John Gardner is an at-large member of the Milwaukee school board. He was elected in 1995 and reelected in 1999.



BROOKINGS
INSTITUTION PRESS

On the Cutting Edge

The Education Gap
Vouchers and Urban Schools
William G. Howell and Paul E. Peterson
cloth, 0-8157-0214-0, **$28.95**

"This is the most important book ever written on the subject of vouchers."
—John E. Brandl, dean, Hubert H. Humphrey Institute of Public Affairs, University of Minnesota

"Love them or hate them, vouchers are on the agenda. Because of the strength of its research design and the daunting financial and administrative hurdles facing anyone who tries to launch another randomized evaluation of vouchers, *The Education Gap* will provide an important intellectual battleground for the debate over vouchers for years to come."
—Alan B. Krueger, Princeton University

Brookings Institution Press
Phone: 800-275-1447 or 202-797-6258
Fax: 202-797-2960 (Order Department)
Available at your local bookstore or order online at
www.brookings.edu

education matters to me

by LAURA KRUG

In the Shadow of Terror

Life returns to not quite normal at Stuyvesant High

It's a bitter-cold morning in New York City as I cross the Stuyvesant Bridge—known to the rest of the world as the TriBeCa Bridge—and hurry toward the school doors, eager to escape the knife-sharp wind. On the threshold, though, I'm stopped mid-step.

"ID card on!" bellows a security guard I've known since I was a freshman. I fumble in my bag, acutely conscious of how late I'm going to be for gym and of how much I detest wearing the blue card. A man who has simply nodded at me every morning of my four years at Stuyvesant High School now needs to see identification before letting me in the school.

Things have definitely changed around here.

PHOTOGRAPH COURTESY OF LAURA KRUG

* * *

Weeks and months have passed. The nightmare that was the morning of September 11 has faded into the background as much as it possibly could have. Stuyvesant High has, accordingly, returned to normal—in most ways, but certainly not all.

It starts with our morning walk down Chambers Street, when we see, down the avenues, the smoke-cloaked masses of debris. During lunch periods, students now are often chased out of the hallways, the gathering places we've always taken for granted. The guards are slowly beginning to relax, but we're still careful about where to sit. We simmered over the fact that until February 1 we weren't allowed to leave the building for lunch, a pre–September 11 privilege that we regarded as a birthright.

Those are just the daily, trivial hassles. The important changes show themselves in social studies classes, where it now takes only the barest spark to begin heated arguments over the United States's forays in the Middle East. Sometimes, a pacifistic argument can be shut down single-handedly with an angry gesture toward the window—in the direction of the still massive mountains of twisted steel and shattered concrete that used to be the World Trade Center. "How can we possibly *not* retaliate," a classmate once snapped, "when there are still 3,000 bodies lying in that pile?" The day's lesson is invariably put on hold whenever the subject of September 11 is brought up. The teachers need the time to teach—even more than usual, as a result of the time that was lost right after the attacks—but they understand that we sometimes need the time to discuss our thoughts and emotions. Whenever anyone hears the noise of a plane nearby, there is a palpable moment of tension. That will be, I think, a hard thing to overcome.

Conflicting reports on air quality come from the Environmental Protection Agency, private air testers, and the city's newspapers. Air-collecting stations are set up around the school; the machinery emits whirring, gurgling noises at all hours. Air testers pace the hallways. When we ask how the air in our classrooms is, they often don't reply. The mystery makes the situation considerably more frightening. A classmate remarked caustically during a discussion once, "What exactly *doesn't* cause cancer?" Perhaps it is a mark of the times that her comment made everyone grin.

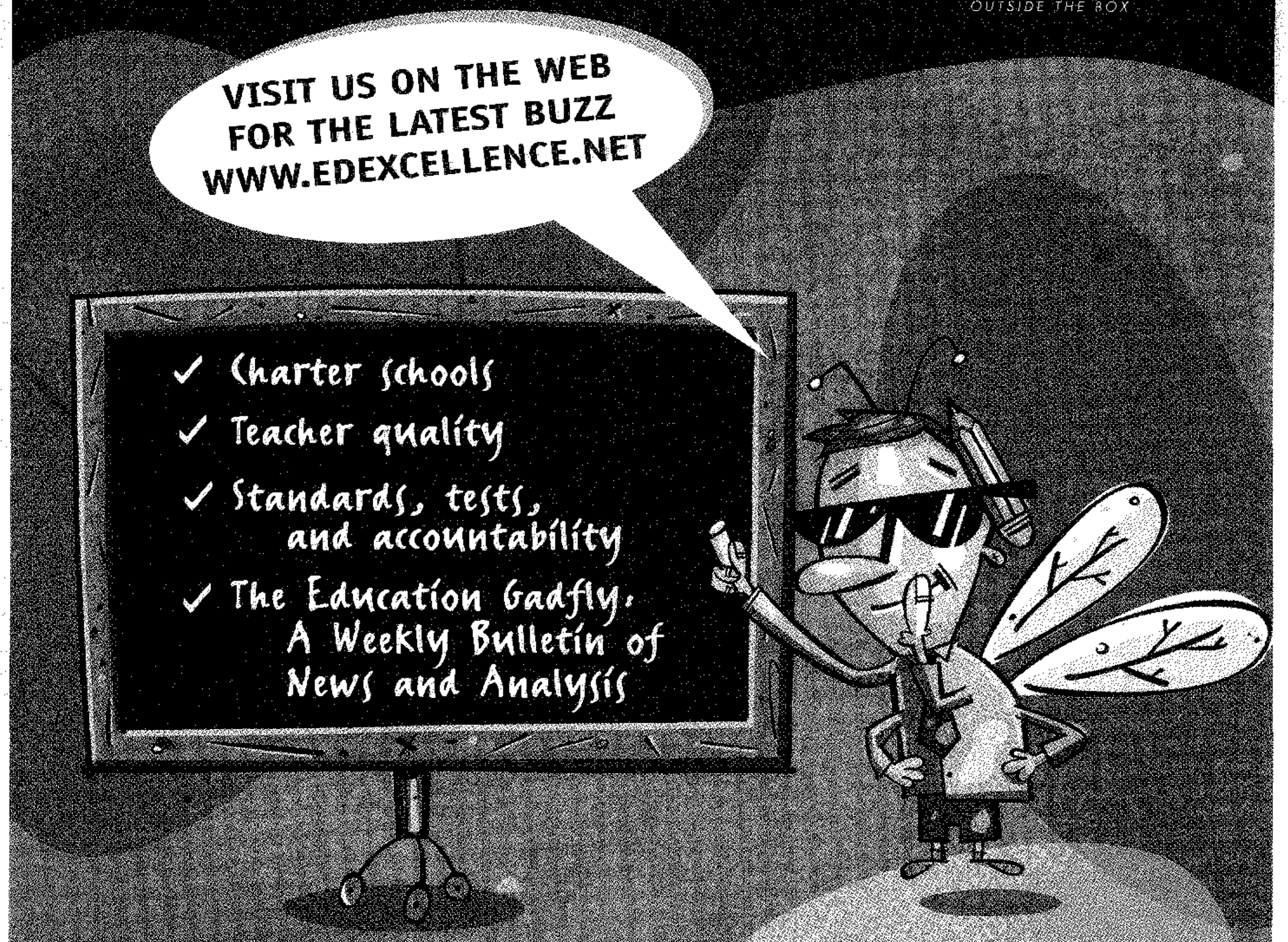
Whenever anyone hears the noise of a plane nearby, there is a palpable moment of tension.

* * *

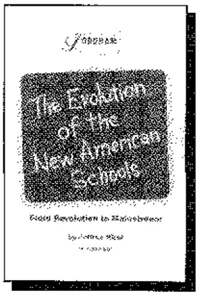
Nevertheless, Stuy is still Stuy. Seniors are receiving their acceptance letters; the hallways are a daily scene of impromptu celebrations and rounds of hugs. Juniors are chugging through junior grind and revving up for their AP exams. The *Spectator* is still printing regularly, sports teams are still playing (even winning sometimes), and friends are still gathering after school in pizzerias. Everyone here was cut to the heart over the events of that terrifying fourth day of school, but we have all helped one another to go on. We share a bond that I believe no other class at Stuyvesant ever has—a bond forged in horror and tragedy, but strengthened by indelible spirit and an immortal sense of togetherness.

–Laura Krug is a senior at Stuyvesant High School in Lower Manhattan and a news editor of the Stuyvesant Spectator.

THOMAS B.
FORDHAM
FOUNDATION
OUTSIDE THE BOX

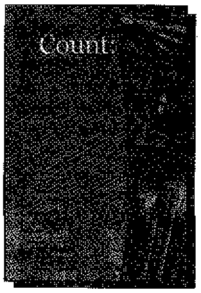

VISIT US ON THE WEB
FOR THE LATEST BUZZ
WWW.EDEXCELLENCE.NET
✓ Charter schools
✓ Teacher quality
✓ Standards, tests, and accountability
✓ The Education Gadfly: A Weekly Bulletin of News and Analysis

RECENT PUBLICATIONS!


The Evolution of the New American Schools

EVOLUTION OF THE NEW AMERICAN SCHOOLS

This start-from-scratch school design initiative was meant to kindle a revolution in American education. This report traces its metamorphosis from bumptious upstart to establishment bulwark.



**MAKING IT COUNT:
A GUIDE TO HIGH IMPACT
EDUCATION PHILANTHROPY**

Practical advice for philanthropists fed up with the status quo and eager to support effective education reforms.

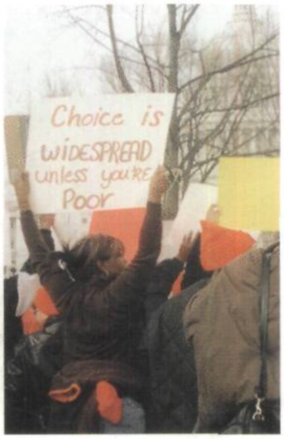
WWW.EDEXCELLENCE.NET

RICHMOND, VIRGINIA
PUBLIC SCHOOL
EDUCATE AT 98
PRIVATE SCHOOL
TEACH AT 100%

Program on Education Policy *and* Governance

HARVARD UNIVERSITY

PARENTS CHOICE


Choice is WIDESPREAD unless you're Poor

CONFERENCES:

Taking Account of Accountability:
Assessing Politics and Policy
June 10–11, 2002

Vouchers and the Supreme Court: What Next?
October 17–18, 2002

PUBLICATIONS:

Charters, Vouchers, and Public Education
Paul E. Peterson and David E. Campbell, editors
Brookings Institution Press, 2002

The Education Gap: Vouchers and Urban Schools
Paul E. Peterson and William G. Howell
with Patrick J. Wolf and David E. Campbell
Brookings Institution Press, 2002

The Program on Education Policy and Governance
Paul E. Peterson, Director
The Taubman Center for State and Local Government
John F. Kennedy School of Government
Harvard University
79 J. F. Kennedy Street, Cambridge, MA 02138
(617) 495-7976
pepg_administrator@ksg.harvard.edu

For more see:
www.ksg.harvard.edu/pepg/

EDUCATION next
A JOURNAL OF OPINION AND RESEARCH
226 Littauer North Yard
1875 Cambridge Street
Cambridge, MA 02138-3001
Tel. (617) 495-2491, Fax (617) 496-1507
Website www.educationnext.org

Nonprofit
U.S. Postage Paid
Permit No. 26
Peterborough, NH
03458

******************AUTO**5-DIGIT 94301
10071 10070 XXXX 1
109 #
Mr. Ron K Unz 267 #
No. 371
555 Bryant St
Palo Alto CA 94301-1704

To subscribe or to submit change of
address information, visit us at
www.educationnext.org;
fax us at (650) 723-8626; email us at
hooverpress@hoover.stanford.edu; or
write to *Education Next* Subscriptions,
Hoover Institution, Stanford University,
Stanford, CA 94305-6010